An Innovative Big Temporal Data Analytics Technique over Real-Life Healthcare Datasets: The F-TBDA Approach

Alfredo Cuzzocrea [1,2,*,†], Geertruida H. de Bock [3], Willemijn J. Maas [3], Selim Soufargi [1], Abderraouf Hafsaoui [1]

[1] iDEA LAB, University of Calabria, Rende, Italy
[2] Department of Computer Science, University of Paris City, Paris, France
[3] Department of Epidemiology, University of Groningen, Groningen, Netherlands

Abstract

In this paper, we introduce and experimentally assess an innovative big data analytics technique for mining and analyzing Quality-of-Life Indicators (QoL) over time among patients with lung cancer and treated with immunotherapy. In more details, given datasets of QoL indicators collected over time, at regular intervals, the F-TBDA technique (Frequency-based Temporal Big Data Analytics) computes temporal relative frequency tables over fixed-time intervals where data of subsequent observations (i.e., intermediate therapy) are compared with the baseline observation (i.e., starting therapy). Then, on the basis of these relative frequency tables, both simple and complex frequency-based big data analytics tools are developed, in order to unveil hidden patterns over cancer patient therapies. Experimental results on top of a real-life dataset nicely complete the theoretical contributions we provide in our research.

Keywords

Big Healthcare Data Analytics, Temporal Relative Frequency Tables, Frequency-Based Big Data Analytics Tools, Clustering, Cancer Patient Data Analytics, Quality of Life Indicators

1. Introduction

Big healthcare data analytics (e.g., [1,2,3]) has become an impressive research area of astonishing relevance during the last years. This relevance is twofold: from a side, it is due to the social impact that these technologies can have over the effectiveness of the healthcare therapies; from another side, at the decision-making level, these technologies can be effectively used for healthcare measures and policies. Recently, a broad set of proposals have appeared, mainly focusing on the rather straightforward idea of applying classical data mining and machine learning to big data but, as shown in authoritative studies (e.g. [4,5]), there are still relevant open issues and research challenges. This is mainly because traditional algorithms are unable to deal with the known characteristics of big data, namely volume, velocity, variety and veracity [6]. Therefore, innovative approaches and solutions need to be devised to effectively and efficiently support big data analytics of healthcare datasets (e.g., [7,8]). The most prominent blueprint to follow in this case is to tailor the design and development phases to the specific case study, trying to achieve a bottom-up methodology (e.g., [21,22])

In line with this goal, the EU H2020 research project QUALITOP - Monitoring multidimensional aspects of QUAlity of Life after cancer ImmunoTherapy: an Open smart digital Platform for personalized prevention and patient management [9] focuses on improving the quality of life of cancer patients treated with immunotherapy. Basically, the project aims at identifying a (broad) set of indicators predictive for Quality of Life (QoL) indicators of cancer patients treated with

SEDB 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

[*] Corresponding author.

[†] This research has been made in the context of the Excellence Chair in Big Data Management and Analytics at University of Paris City, Paris, France.

alfredo.cuzzocrea@unical.it (A. Cuzzocrea); g.h.de.bock@umcg.nl (G.H. de Bock); w.j.maas@umcg.nl (W.J. Maas); selim.soufargi@unical.it (S. Soufargi); ahafsaoui.idealab.unical@gmail.com (A. Hafsaoui)

0000-0002-7104-6415 (A. Cuzzocrea); 0000-0003-3104-4471 (G.H. de Bock); 0000-0002-0792-7090 (W.J. Maas); 0009-0000-5476-9403 (S. Soufargi); 0000-0002-2364-172X (A. Hafsaoui)

immunotherapy. and processing, mining and analyzing these indicators. The ultimate goal is to come up with active guidelines that contribute to improving patients' well-being at every level, including the physical level, the social/family level, the emotional level and the functional level (e.g., [10]). Within this context, this paper focuses attention on an innovative big data analytics technique called Frequency-based Temporal Big Data Analytics (F-TBDA), with the main goal of discovering hidden temporal patterns from periodic consecutive QoL indicator observations over time. In more detail, given a dataset of baseline QoL indicator observations, denoted by B, and T fixed-time subsequent follow-up QoL indicator observations, denoted by $F = \{F_0, F_1, \dots, F_{T-1}\}$, the F-TBDA technique computes a set of temporal relative frequency tables over fixed time intervals, denoted by $FT = \{FT_0, FT_1, \dots, FT_{T-1}\}$, such that each element of each frequency table $FT_k[i,j] \in FT$ stores the (relative) frequency of the difference between the follow-up QoL indicator value in $F_k[i,j] \in F$ and the baseline QoL indicator value in $B[i,j]$. Clearly, the difference scores capture the evolution of that particular QoL indicator from the baseline case to the k -time case. Based on this temporal pattern mining approach as done with F-TBDA, the proposed framework implements several frequency-based big data analytics tools, to reveal hidden temporal patterns about the effects of various patient, disease and treatment therapies with respect to QoL.

Another strong contribution of this paper consists of the experimental assessment and analysis of F-TBDA on top of a real-life dataset of lung-cancer patients treated with immunotherapy, with QoL assessments at baseline up to four distinct follow-up assessments, namely at three, six, twelve and eighteen months. Experimental results on top of a real-life dataset nicely complete the theoretical contributions we provide in our research. The full version of this contribution appears in [43].

2. Related Work

In this Section, we provide an overview on research proposals that are close to our work.

In [11], authors describe a technique useful to discover trends in healthcare data that is a conjunction of association rule mining with model-based recursive partitioning [12]. Indeed, the technique allows the discovery of temporal trends through association rules based on a user-defined diagnosis. Mined rules set the potential trends among the patients and the model-based recursive partitioning inspects how strong is the trend across time while accounting for demographic attributes such as the age of the patients. Diagnosis are used as input for the Apriori algorithm [13] to generate rules, and the highest confidence results are adopted to generate data to fed to the model-based recursive partitioning algorithm in order to yield the temporal trends. Specifically, Gender and Age are the two attributes used as partitioning variables for the built regression tree that results from applying a model-based recursive partitioning, while the admission month or quarter was the regressor. Finally, the leaves of the built tree constitute trends related to the patient admission months. Experimental results are presented for Hyperlipidemia as well as for Hypertension diagnoses based on data from the NIS dataset [14].

In [15], authors propose a methodology to mine temporal association rules from hybrid healthcare event data using a variation of the Apriori algorithm [16]. Indeed, an administrative and clinical data mesh-up is generated and constitutes the input data for the mentioned mining algorithm. Clinical data are then pre-processed through a knowledge-based Temporal Abstraction Technique to ensure the time-series related to measurements taken on a particular patient regarding a specific attribute (for example, their glucose levels or concomitant cancers) are turned into qualitative descriptions that can be integrated into the administrative data. The final goal of the process is to find physiological aspects of the patients that show a significant temporal association with the drug prescriptions. For instance, a pertinent result from applying the algorithm on the pre-processed hybrid event data would be that overweight patients having high Glycaemia and excessively high HbA1c would be prescribed Platelet aggregation inhibitors with certain confidence value and within a time-window which default to six months.

Authors in [17] describe a methodology to discover temporal phenotypes through Lines Of Therapy (LOT) in cancer patients, namely lung-cancer and melanoma patients. Indeed, identifying temporal phenotypes in cancer patients undergoing a therapy would assist in spotting observable traits related to disease progression or reaction to drug administration in a timely manner that eventually would provide insights into adjusting therapies across specific patient populations and hence enabling precision medicine. Because treatments are selected based on phenotypes (morphological, biochemical, physiological, behavioral, and so forth), it is important to detect patient phenotypes over the course of their treatment period. In this sense, the current work proposes to extract these phenotypes by k-means clustering [18] LOT of melanoma and lung-cancer patients. The LOT data are in turn extracted using an algorithm that takes as input insurance claims raw data as well as Electronic Health Records (EHR) data and transform them to yield in a first step data structured in terms of patient cohorts (based on demographic values) and in a second step systemic anticancer therapy claims data. Resulting clusters would most importantly differentiate slow progression vs fast progression phenotypes across the patient cohorts. The data are finally used to plot Sankey diagrams related to phenotyping clustering across time.

Finally, in [19] a Web-based platform is proposed. This platform adheres to the Model-View-Controller (MVC) architectural pattern: the model uses data sharding to store data in a NoSQL database via healthcare tailored tree-shaped set of documents (equivalent of tables in SQL), which is named PareMeDocs. This solution ensures highly performant search and filter querying. Data sharding is either system-based, or user-defined. In the case of user-defined sharding, meaningful shards of data could be created for storage through, for instance, specific attributes useful to speed-up underlying data querying (e.g., data could be shared by seasons of the year by gathering and storing same-season records in the same shard, and that based on a date attribute). Eventually, authors aim at drastically increase search/filter queries over such data by leveraging Map-Reduce over sharded data compiled in a PareMeDocs tree (map functions identify target shards of data for the query and the reduce aggregate those data accordingly). Application interface provides a timely-display of the records of the NHIRD dataset (~1B records) [20]. For example, for a specific patient ID or given their birthday or sex, a timeline depiction is displayed to the user in order to inspect each record according to specific attributes. The setting offers a wide range of statistical (distributions, aggregation, etc.) and/or chronological charts over queried sub-populations.

3. Dataset Description and Medical Analysis Goals

In this Section, we provide the description of the real-life QoL dataset and the specific medical goals related to its definition, creation, and population. We named this dataset as QUALITOP DataSet (QDS). Data stored in the QDS dataset were collected within the context of QUALITOP [23], at the University Medical Centre Groningen (UMCG) and Treant hospital, The Netherlands, between May 2021 and October 2023. The infrastructure of the UMCG data-biobank OncoLifeS was used to collect these data [24]. Patients included are aged ≥ 18 years at the time of signing informed consent, diagnosed with lung cancer and treated with Immunotherapy, Immune Checkpoint Inhibitors (ICIs). Patients received monotherapy or in combination with other cancer treatments, for example, chemotherapy of radiotherapy. Inclusion was possible before the start of the second cycle of treatment. Patients were excluded when they were pregnant, under guardianship or refused to sign informed consent.

To assess QoL indicators, we used the Functional Assessment of Cancer Therapy – General (FACT-G) methodology, which is specifically developed for patients receiving cancer treatment and is widely used since the nineties [25,26]. FACT-G consists of 27 questions divided in four domains: physical well-being, social/family well-being, emotional well-being, and functional well-being. For our patients, the authorized Dutch version was used. FACT-G was collected at baseline, start of immunotherapy, at 3-months follow-up, 6-months follow-up, 12-months follow-up, and 18-months follow-up. Patients were asked to fill in the questionnaires, regardless of continued treatment. Clinical data was collected from the electronic patient report and collected and managed in Research Electronic Data Capture (REDCap) [27], which is a secure, web-based platform designed to support data for research studies. Furthermore, REDCap sent automatically the FACT-G questionnaire at baseline and follow-up moments to patients by e-mail. Some patients requested to receive a paper version of the FACT-G at home.

The purpose of this analysis is to identify patterns in quality-of-life over time and to link these to patients, diseases and treatment characteristics. Indeed, quality-of-life is the perceived quality of an individual's daily life, that is, an assessment of their well-being or lack thereof. In the specific healthcare setting, health-related quality-of-life is an assessment of how the individual's well-being may be affected over time by a disease, disability or disorder [28]. Examples of quality-of-life characteristics include: age, gender, disease stage, type of treatment, etc. Based on the results found, it is possible to determine which patients are more or less likely to benefit from immunotherapy before the start of treatment. By including the patient perspective in research, the patient can be better informed before starting treatment. This will lead to better decision making.

On a wider perspective, these methodologies are useful to support advanced analytics for healthcare policies at a greater level (e.g., regional level, national level, and so forth – e.g. [29]). In fact, decision-making processes of a certain region can scale from the hospital level to the regional management level, by devising innovative cooperative paradigms.

4. The F-TBDA Technique: Frequency-Based Temporal Big Data Analytics

In this Section, we describe in details the main contribution of this research, i.e., the F-TBDA technique.

First, given: (i) a medical cancer therapy temporal observation dataset D having P records, (ii) a set of baseline QoL attributes $\{q_0, q_1, \dots, q_{N-1}\}$, selected from D, denoted as Q, (iii) the corresponding R follow-up QoL attribute groups of the k-months observation $\{f_{0,kM}, f_{1,kM}, \dots, f_{N-1,kM}\}$, denoted as F_{kM}, we derive $R+1$ input corresponding datasets, namely $Q \cup \{F_{k_0M}, F_{k_1M}, F_{k_2M}, \dots, F_{k_{R-1}M}\}$. For instance, if $R = 4$ and $k \in \{3,6,12,18\}$, we derive 5 input datasets, like depicted in Figure 1.

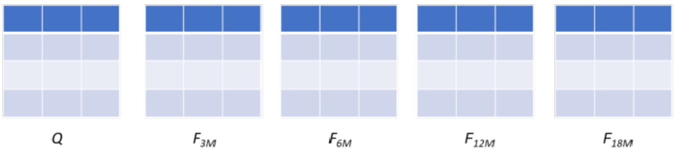


Figure 1: Derived Datasets for the Case $R = 4$ and $k = 3$

In Figure 1, (i) Q stores P records having $\{q_0, q_1, \dots, q_{N-1}\}$ as attributes; (ii) F_{3M} stores P records having $\{f_{0,3M}, f_{1,3M}, \dots, f_{N-1,3M}\}$ as attributes; (iii) F_{6M} stores P records having $\{f_{0,6M}, f_{1,6M}, \dots, f_{N-1,6M}\}$ as attributes; (iv) F_{12M} stores P records having $\{f_{0,12M}, f_{1,12M}, \dots, f_{N-1,12M}\}$ as attributes; (v) F_{18M} stores P records having $\{f_{0,18M}, f_{1,18M}, \dots, f_{N-1,18M}\}$ as attributes.

In order to capture the variations of sub-sequent QoL follow-ups with respect to the QoL baselines, we carefully build R $P \times N$ QoL tables, denoted as QOL_{kM}, such that each element $QOL_{kM}[i,j]$, where $k \in \{k_0, k_1, \dots, k_{R-1}\}$, is defined as follows:

$$QOL_{kM}[i,j] = |F_{kM}[f_{j,kM}][i] - Q[q_j][i]|, \quad \forall k \in \{k_0, k_1, \dots, k_{R-1}\} \tag{1}$$

where, given a table T in the set $Q \cup F_{kM}$, where $k \in \{k_0, k_1, \dots, k_{R-1}\}$, $T[a_j][i]$ denotes the i-th value of the attribute $a_j \in T$. Still in the previous case study, we derive 4 QoL tables, like depicted in Figure 2.

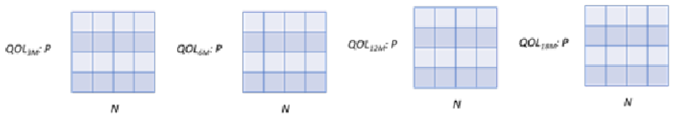


Figure 2: Derived QOL Tables for the Running Case Study

Then, for each QoL table QOL_{kM}, we compute a proper $P \times N$ relative frequency table, denoted by $FT_{Q,kM}$, such that each element $FT_{Q,kM}[i,j]$, where $k \in \{k_0, k_1, \dots, k_{R-1}\}$, is defined as follows:

$$FT_{Q,kM}[i,j] = \frac{QOL_{Q,kM}[i,j]}{\sum_{i=0}^{N} QOL_{Q,kM}[i,j]}, \forall k \in \{k_0, k_1, \dots, k_{R-1}\} \tag{2}$$

By applying this methodology to all the frequency tables of the model, we finally obtain a temporal relative frequency table stream, on top of which several frequency-based big data analytics tools can be applied in order to discovery hidden temporal patterns from QoL data. Still in the previous case study, we derive 4 QoL frequency tables, as depicted in Figure 3.

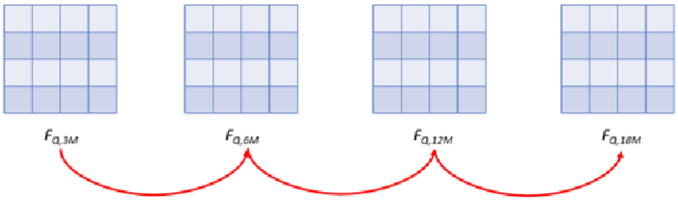


Figure 3: Temporal Relative Frequency Table Stream for the Running Case Study

Based on this so-defined analytical model, we introduce two different frequency-based big data analytics tools. The first one, named as Temporal Variation of Relative Frequencies (TVRF), is oriented to study the evolution of specific QoL attributes, by inspecting its relative frequency variation over time. The second one, named as Clustering Similarity Analysis over Temporal Frequency Tables (CSATFT), is oriented to study the evolution of all the QoL attributes, by inspecting the similarities of different clusterings computed over frequency tables across time. In the following, we focus on these tools.

As regards the TVRF analytics, given a QoL attribute $q_i \in FT_{Q,kM}$, such that $q_i \in \{q_0, q_1, \dots, q_{N-1}\}$ and $k \in \{k_0, k_1, \dots, k_{R-1}\}$, we report the percentage variation of the relative frequency of q_i across time, by accessing the corresponding q_i attribute value across different frequency tables. Figure 4 shows a simple yet illustrative example of the TVRF analytics.

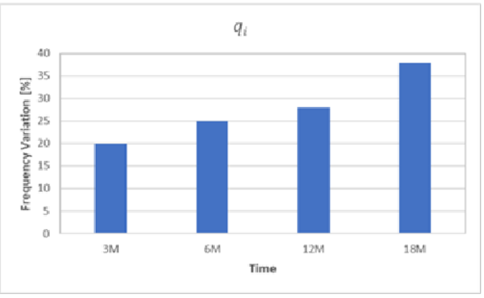


Figure 4: TVRF Analytics Example

Formally, every variation in the relative frequency plot of $q_i = FT_{Q,kM}[i]$ across time, such that $i \in \{0,1,\dots,N-1\}$ and $k \in \{k_0, k_1, \dots, k_{R-1}\}$, denoted by $\widetilde{q_{i,l}}$, is obtained as follows:

$$\widetilde{q_{i,l}} = \frac{1}{P}\sum\nolimits_{i=0}^{P-1} FT_{Q,kM}[i], \qquad (3)$$
$$\forall i \in \{0, \dots, N-1\}, \forall k \in \{k_0, k_1, \dots, k_{R-1}\}$$

As regards the CSATFT analytics, given a frequency $FT_{Q,kM}$, such that $k \in \{k_0, k_1, \dots, k_{R-1}\}$, we interpret every row of $FT_{Q,kM}$ as a data object, and we apply a clustering algorithm to the so-derived data domain of $FT_{Q,kM}$. In particular, among all the alternatives, we selected k-means clustering [18], following the guidelines of several authoritative studies in the area (e.g., [30]). Figure 5 shows a simple yet illustrative example of the clustering for the toy frequency table $FT_{Q,3M}$.

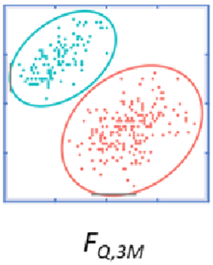


Figure 5: Clustering Example for the Toy Frequency Table $FT_{Q,3M}$

By applying this methodology on top of all the stream of temporal relative frequency tables, we obtain accordingly a stream of clustering. Figure 6 shows an illustrative example when considering the frequency tables of Figure 3.

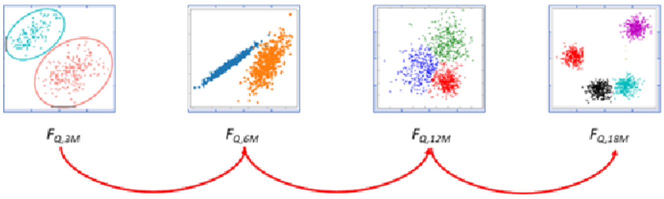


Figure 6: Clustering over the Temporal Relative Frequency Table Stream of Figure 3

Based on the so-derived stream of clusterings we compute a suitable clustering similarity measure among different clusterings over time. Among all the alternatives, by following the considerations of other similar studies in the field (e.g., [31]), we selected the Cosine similarity measure [32], by applying a specific variation suitable to our case study.

Given two clusterings Φ_i_i and Φ_j, and given two different clusters C_i and C_j, such that $C_i \in \Phi_i$ and $C_j \in \Phi_j$, respectively, the Cosine similarity between C_i and C_j, denoted by $S(C_i, C_j)$, is defined as follows:

$$S(C_i, C_j) = 1 + \frac{1}{\Delta(C_i, C_j)} \tag{4}$$

where:

$$\Delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i y \in C_j} \delta(x, y) \tag{5}$$

$\delta(x, y)$ is the distance between two data points x and y, which is defined as in the next. Let $F_x(i)$ be the i -th numerical feature and $D_x(i)$ be the i -th categorical feature of a data point x. Then, $\delta(x, y)$ is defined as follows:

$$\delta(x, y) = \sum_i \gamma_i |F_x(i) - F_y(i)|^p + \eta \sum_j \|D_x(j) - D_y(j)\|_2^2 \tag{6}$$

such that γ_i, p and η are tuning parameters to be chosen on the basis of specific distributions of the target dataset [32].

Based on the definition of the Cosine similarity of two clusters C_i and C_j, the Cosine similarity between two clusterings Φ_1 and Φ_2, $S(\Phi_1, \Phi_2)$, is finally defined as follows:

$$S(\Phi_1, \Phi_2) = max\{S(C_i, C_j), S(C_i, C_k)\}, \forall i \in \Phi_1 \; \forall j, k \in \Phi_2 \tag{7}$$

Figure 7 shows a simple yet illustrative example of the CSATFT analytics, where the clustering similarity plot between, the toy frequency tables $FT_{Q,3M}$, and $FT_{Q,6M}$, $FT_{Q,12M}$ and $FT_{Q,18M}$, respectively, is shown.

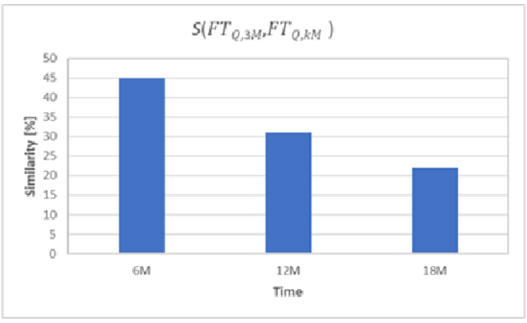


Figure 7: CSATFT Analytics Example

5. Experimental Evaluation and Analysis

In this Section, we provide a comprehensive experimental evaluation and analysis of the big data analytics performance of the F-TBDA technique, by performing several experimental tests over the real-life QDS dataset.

As regards the proper experimental engineering we performed in our campaign, in order to deal with the high-dimensionality of the QDS dataset, we first partitioned the dataset in six distinct datasets, which we list in the following: (i) QDS_BASELINE, which stores the starting values of the QoL attributes at the beginning of the immunotherapy; (ii) QDS_3MONTHS_FUP, which stores the values of the QoL attributes after three months from the start of immunotherapy, i.e. the three-month follow-up; (iii) QDS_6MONTHS_FUP, which stores the values of the QoL attributes after six months from the start of immunotherapy, i.e. the six-month follow-up; (iv) QDS_12MONTHS_FUP, which stores the values of the QoL attributes after twelve months from the start of immunotherapy, i.e. the twelve-month follow-up; (v) QDS_18MONTHS_FUP, which stores the values of the QoL attributes after eighteen months from the start of immunotherapy, i.e. the eighteen-month follow-up; (vi) QDS_CLINICAL, which stores the values of the clinical assessments of patients associated to the QoL attributes.

In our experimental analysis, since we specifically focus on QoL evolutions, we consider the first five datasets on top of the overall six datasets. The correlation of QoL evolutions with clinical assessment outcomes is thus left to future research efforts in the field.

After this phase, among the high number of QoL attributes, we selected a sub-set of them, in order to perform our experimental study. In particular, being inspired by well-understood feature selection analysis (e.g., [33]) the following domains have been selected, in accordance with the FACT-G methodology (see Section 3): (i) physical well-being: this domain of attributes models aspects related to the physical well-being of cancer patients; (ii) social/family well-being: this domain of attributes models aspects related to the social and family well-being of cancer patients; (iii) emotional well-being: this domain of attributes models aspects related to the emotional well-being of cancer patients; (iv) functional well-being: this domain of attributes models aspects related to the functional well-being of cancer patients.

It should be noted that these attribute domains well capture the overall QoL process for the both the goals of effectiveness and efficiency, according to several recent studies (e.g., [34]). However, due to space limit, we focus our experimental analysis on the physical well-being domain, which is an authoritative representative of the quality of life of cancer patients. The selected QoL physical well-being attributes are reported in Table 1. In particular, Table 1 shows the meaning of each QoL attribute and its possible range of numerical encodings from 1 to 5, with the associated meaning.

Table 1

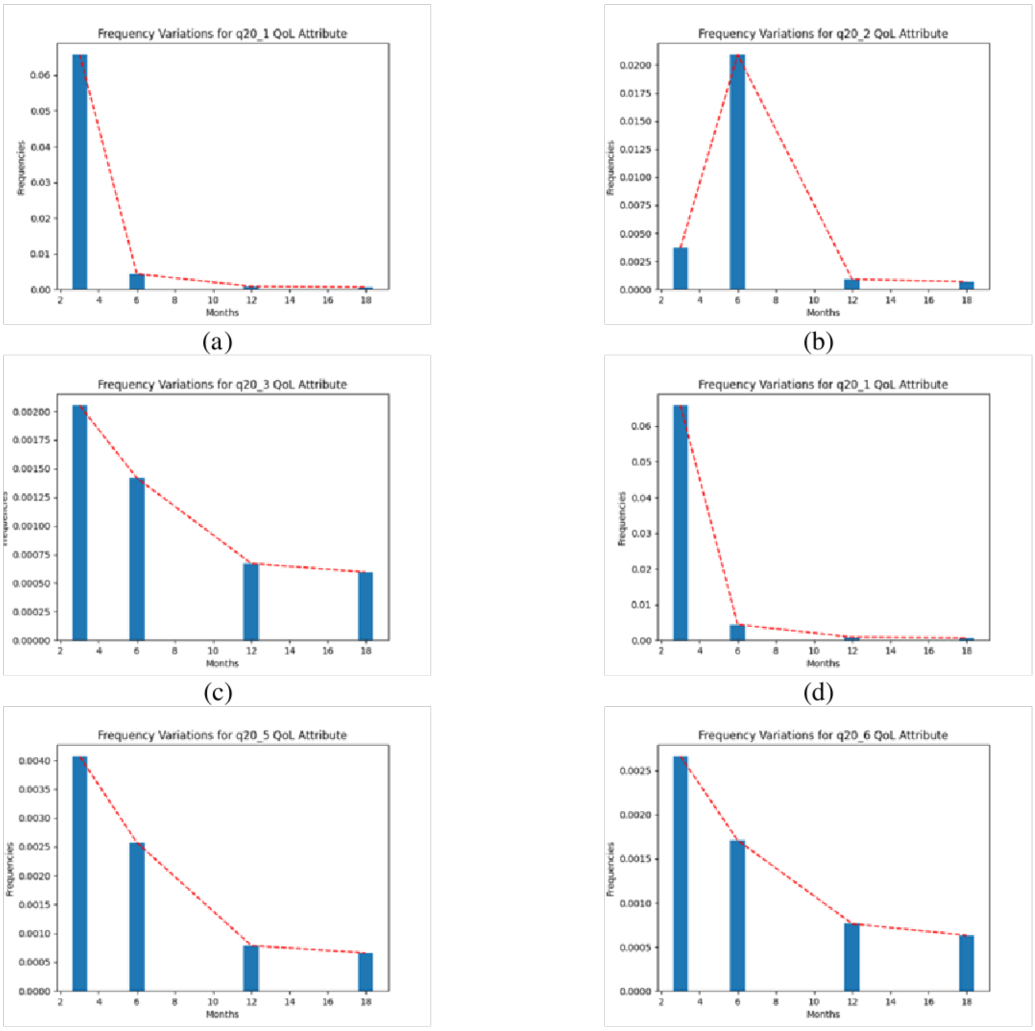
QOL Attributes of the Physical Well-Being Domain

Physical Well-Being Domain	
QoL Attribute	**Description**
q_{20_1}	“I have a lack of energy” – {1: Not at all | 2: A little bit | 3: Somewhat | 4: Quite a lot | 5: Very much}
q_{20_2}	“I feel nausea” – {1: Not at all | 2: A little bit | 3: Somewhat | 4: To a fairly high degree | 5: To a very high degree}
q_{20_3}	“Because of my physical condition, I have trouble meeting the needs of my family” – {1: Not at all | 2: A little bit | 3: Somewhat | 4: To a fairly high degree | 5: To a very high degree}
q_{20_4}	“I have pain” – {1: Not at all | 2: A little bit | 3: Somewhat | 4: To a fairly high degree | 5: To a very high degree}
q_{20_5}	“I am bothered by the side effects of the treatment” – {1: Not at all | 2: A little bit | 3: Somewhat | 4: To a fairly high | 5: To a very high degree}
q_{20_6}	“I feel ill” – {1: Not at all | 2: A little bit | 3: Somewhat | 4: To a fairly high degree | 5: To a very high degree}

It should be noted that, using encodings to model the possible values of the QoL attributes, follows common Machine Learning (ML) practices in data analytics (e.g., [35]).

According to what is described in Section 4, we organized our experimental evaluation by means of the main big data analytics tools (i.e., TVRF and CSATFT), which mainly focus the attention on stressing the accuracy of our big data analytics tools, while the performance assessment is left as future work. Due to space limitation, here we focus on the TVRF analytics only.

For every QoL physical well-being attribute, we thus derived the corresponding TVRF plot, in order to capture the frequency variations over time, across the different follow-up values of the QoL attributes. Figure 8 reports the TVRF plots for the QoL attributes$\{q_{20_1}, q_{20_2}, q_{20_3}, q_{20_4}, q_{20_5}, q_{20_6}\}$.



(a) (b) (c) (d) (e) (f)

Figure 8: TVRF q_{20_1} (*a*) – TVRF q_{20_2} (*b*) – TVRF q_{20_3} (*c*) – TVRF q_{20_4} (*d*) – TVRF q_{20_5} (*e*) – TVRF q_{20_6} (*f*)

From the analysis of Figure 8, it clearly follows that, for almost all the QoL physical well-being attributes, after the 3-months follow-up therapy, variations stabilize towards stable (and lower) values. This shows the effectiveness of the therapy itself. On the other hand, q_{20_2} variations show a singular trend. For this certain attribute, “I have nausea”, more research is needed to explain why at 6-months follow-up there is more variation compared to the 3-months follow-up.

6. Conclusions and Future Work

Starting from the real-life experiences achieved with the context of the QUALITOP EU H2020 research project, in this paper we have introduced and experimentally assessed F-TBDA, an innovative frequency-based temporal big data analytics technique aimed at discovering hidden temporal patterns from quality-of-life indicators of cancer patients. Our technique makes use of advanced methodologies developed on top of frequency tables built from baseline and follow-up QoL data, mostly based on clustering and clustering similarity analysis. Retrieved results have confirmed the benefits of F-TBDA in discovering hidden temporal patterns from quality-of-life indicators of cancer patients, and its clear suitability for being extended with novel artificial intelligence methods.

Future work is mainly oriented towards extending our actual frequency-based temporal big data analytics framework with innovative characteristics of the emerging big data trend, such as: (i) multidimensionality of data (e.g., [36,37]); (ii) privacy of data (e.g., [38,39]); (iii) data approximation metaphors (e.g., [40,41]); (iv) data processing performance (e.g., [42]).

Acknowledgment

This research has been supported by the EU H2020 QUALITOP research project, H2020 Project ID: 875171.

References

[1] S. Dash, S.K. Shakyawar, M. Sharma, S. Kaushik, "Big Data in Healthcare: Management, Analysis and Future Prospects". J. Big Data 6, art. 54, 2019.

[2] K.M. Batko, A. Slezak, "The Use of Big Data Analytics in Healthcare". J. Big Data 9(1), art. 3, 2022.

[3] Z. Fei, Y. Ryeznik, O. Sverdlov, C.-W. Tan, W.K. Wong, "An Overview of Healthcare Data Analytics with Applications to the COVID-19 Pandemic". IEEE Trans. Big Data 8(6), pp. 1463-1480, 2022.

[4] W. Li, Y. Chai, F. Khan, S.R.U. Jan, S. Verma, V.G. Menon, Kavita, X. Li, "A Comprehensive Survey on Machine Learning-Based Big Data Analytics for IoT-Enabled Smart Healthcare System". Mob. Networks Appl. 26(1), pp. 234-252, 2021.

[5] D.A. Pustokhin, I.V. Pustokhina, P. Rani, V. Kansal, M. Elhoseny, G.P. Joshi, K. Shankar, "Optimal Deep Learning Approaches and Healthcare Big Data Analytics for Mobile Networks Toward 5G". Comput. Electr. Eng. 95, art. 107376, 2021.

[6] D. Laney, "3D Data Management: Controlling Data Volume, Velocity and Variety". META Group Research Note, 6(70), 2001.

[7] C.-K. Ngan, Y.-F. Paat, R. Green, "HDSS: A Healthcare Decision Support System on Combining Domain Knowledge and Data Analytics for Predicting Potential Risk of Mental Health". Int. J. Appl. Decis. Sci. 15(4), pp. 465-491, 2022.

[8] K. Zheng, S. Cai, H.R. Chua, M. Herschel, M. Zhang, B.C. Ooi, "DyHealth: Making Neural Networks Dynamic for Effective Healthcare Analytics". Proc. VLDB Endow. 15(12), pp. 3445-3458, 2022.

[9] The QUALITOP Research Project, Available: https://h2020qualitop.liris.cnrs.fr/wordpress/index.php/project/

[10] E. Klecun, "Transforming Healthcare: Policy Discourses of IT and Patient-Centred Care". Eur. J. Inf. Syst. 25(1), pp. 64-76, 2016.

[11] G. Hrovat, G. Stiglic, P. Kokol, M. Ojsteršek, "Contrasting Temporal Trend Discovery for Large Healthcare Databases". Comput Methods Programs Biomed. 113(1), pp. 251-257, 2014.

[12] A. Zeileis, T. Hothorn, K. Hornik, "Model-Based Recursive Partitioning". Journal of Computational and Graphical Statistics 17(2), pp. 492-514, 2008.

[13] R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules". Proceedings of the 20th International Conference on Very Large DataBases, pp. 487-499, 1994.

[14] Agency for Healthcare Research and Quality, HCUP Nation wide Inpatient Sample (NIS), Healthcare Costand Utilization Project (HCUP). Rockville, MD, 2009.

[15] S. Concaro, L. Sacchi, C. Cerra, P. Fratino, R. Bellazzi, "Mining Healthcare Data with Temporal Association Rules: Improvements and Assessment for a Practical Use". Proceedings of the 12th Artificial Intelligence in Medicine, pp. 16-25, 2009.

[16] S. Concaro, L. Sacchi, C. Cerra, P. Fratino, R. Bellazzi, "Temporal Data Mining for the Analysis of Administrative Healthcare Data". Proceedings of IDAMAP Workshop. pp. 75-80, 2008.

[17] W. Meng, W. Ou, S. Chandwani, X. Chen, W. Black, Z. Cai, "Temporal Phenotyping by Mining Healthcare Data to Derive Lines of Therapy for Cancer". J Biomed Inform. 100, 2019.

[18] J.B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 281-297, 1967.

[19] C.C.-H. Lin, L.-C. Huang, S.-C.T. Chou, C.-H. Liu, H.-F. Cheng, I-J. Chiang, "Temporal Event Tracing on Big Healthcare Data Analytics". Proceedings of IEEE Big Data Congress 2014, pp. 281-287, 2014.

[20] National Health Insurance Research Database, Available: http://nhird.nhri.org.tw/en/index.htm

[21] F. Movahedi, Y. Zhang, R. Padman, J.F. Antaki, "Mining Temporal Patterns from Sequential Healthcare Data". Proceedings of 2018 IEEE International Conference on Healthcare Informatics, pp. 461-462, 2018.

[22] C. Meaney, M. Escobar, R. Moineddin, T.A. Stukel, S. Kalia, B. Aliarzadeh, T. Chen, B. O'Neill, M. Greiver, "Non-Negative Matrix Factorization Temporal Topic Models and Clinical Text Data Identify COVID-19 Pandemic Effects on Primary Healthcare and Community Health in Toronto, Canada". J. Biomed. Informatics 128, 2022.

[23] P.C. Vinke, M. Combalia, G.H. de Bock, et al., "Monitoring Multidimensional Aspects of Quality of Life after Cancer Immunotherapy: Protocol for the International Multicentre, Observational QUALITOP Cohort Study". BMJ Open 13(4), art. e069090, 2023.

[24] G. Sidorenkov, J. Nagel, C. Meijer, J.J. Duker, H.J.M. Groen, G.B. Halmos, M.H.M. Oonk, R.J. Oostergo, B. van der Vegt, M.J.H. Witjes, M. Nijland, K. Havenga, J.H. Maduro, J.A. Gietema, G.H. de Bock, "The Oncolifes Data-Biobank for Oncology: A Comprehensive Repository of Clinical Data, Biological Samples, and the Patient's Perspective". Journal of Translational Medicine 17, 2019.

[25] A.E. Bonomi, D.F. Cella, E.A. Hahn, et al., "Multilingual Translation of the Functional Assessment of Cancer Therapy (FACT) Quality of Life Measurement System". Qual Life Res. 5, pp. 309–320, 1996.

[26] D. Cella, L. Hernandez, A.E. Bonomi, et al., "Spanish Language Translation and Initial Validation of the Functional Assessment of Cancer Therapy Quality-Of-Life Instrument". Med Care. 36(9), pp. 1407-1418, 1998.

[27] P.A. Harris, R. Taylor, B.L. Minor, et al., “The REDCap Consortium: Building an International Community of Software Platform Partners”. J Biomed Inform. 95, 2019.

[28] C.H. Hennessy, D.G. Moriarty, M.M. Zack, P.A. Scherr, R. Brackbill, “Measuring Health-Related Quality of Life for Public Health Surveillance”. Public Health Rep. 109(5), pp. 665-672, 1994.

[29] K. Chao, M.N.I. Sarker, I. Ali, R.B.R. Firdaus, A. Azman, M.M. Shaed, “Big Data-Driven Public Health Policy Making: Potential for the Healthcare Industry”. Heliyon 9(9), 2023.

[30] S. Soheily-Khah, A.D. Chouakria, É. Gaussier, “Generalized k-Means-Based Clustering for Temporal Data under Weighted and Kernel Time Warp”. Pattern Recognit. Lett. 75, pp. 63-69, 2016.

[31] J. Zheng, D. Chen, H. Hu, “Boundary Adjusted Network Based on Cosine Similarity for Temporal Action Proposal Generation”. Neural Process. Lett. 53(4), pp. 2813-2828, 2021.

[32] A. Singhal, “Modern Information Retrieval: A Brief Overview”. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24(4), pp. 35-43, 2001.

[33] B. Ghojogh, M.N. Samad, S.A. Mashhadi, T. Kapoor, W. Ali, F. Karray, M. Crowley, “Feature Selection and Feature Extraction in Pattern Analysis: A Literature Review”. arXiv preprint, arXiv:1905.02845, 2019.

[34] M. Savic, V. Kurbalija, M. Ilic, M. Ivanovic, D. Jakovetic, A. Valachis, S. Autexier, J. Rust, T. Kosmidis, “The Application of Machine Learning Techniques in Prediction of Quality of Life Features for Cancer Patients”. Comput. Sci. Inf. Syst. 20(1), pp. 381-404, 2023.

[35] I. Lopez-Arevalo, E. Aldana-Bobadilla, A. Molina-Villegas, H. Galeana-Zapién, V. Muñiz-Sanchez, S. Gausin-Valle, “A Memory-Efficient Encoding Method for Processing Mixed-Type Data on Machine Learning”. Entropy 22(12), art. 1391, 2020.

[36] A. Cuzzocrea, “Analytics over Big Data: Exploring the Convergence of DataWarehousing, OLAP and Data-Intensive Cloud Infrastructures”. Proceedings of 2013 IEEE COMPSAC, pp- 481-483, 2013.

[37] A. Cuzzocrea, Jérôme Darmont, Hadj Mahboubi, “Fragmenting Very Large XML Data Warehouses via K-Means Clustering Algorithm”. Int. J. Bus. Intell. Data Min. 4(3/4), pp. 301-328, 2009.

[38] A. Cuzzocrea, Domenico Saccà, “Balancing Accuracy and Privacy of OLAP Aggregations on Data Cubes”. Proceedings of 13th ACM DOLAP, pp. 93-98, 2010.

[39] A. Cuzzocrea, Vincenzo Russo, “Privacy Preserving OLAP and OLAP Security”. Encyclopedia of Data Warehousing and Mining 2009, pp. 1575-1581, 2009.

[40] A. Cuzzocrea, “Providing Probabilistically-Bounded Approximate Answers to Non-Holistic Aggregate Range Queries in OLAP”. Proceedings of 8th ACM DOLAP, pp. 97-106, 2005.

[41] A. Cuzzocrea, Ugo Matrangolo, “Analytical Synopses for Approximate Query Answering in OLAP Environments”. Proceedings of 15th DEXA, pp. 359-370, 2004.

[42] B. Yu, A. Cuzzocrea, D.H. Jeong, S. Maydebura, “On Managing Very Large Sensor-Network Data Using Bigtable”. Proceedings of 2012 IEEE CCGRID, pp. 918-922, 2012.

[43] Alfredo Cuzzocrea, Geertruida H. de Bock, Willemijn J. Maas, Selim Soufargi, ”F-TBDA: A Frequency-Based Temporal Big Data Analytics Technique for Mining and Analyzing Quality-Of-Life Indicators of Cancer Patients”. Proceedings of IEEE Big Data 2023, pp. 5197-5205, 2023