Medical Informatics in a United and Healthy Europe
K.-P. Adlassnig et al. (Eds.)
IOS Press, 2009

doi:10.3233/978-1-60750-044-5-579

Knowledge Discovery about Quality of Life Changes of Spinal Cord Injury Patients: Clustering Based on Rules by States

Karina GIBERT [a,1], Alejandro GARCÍA-RUDOLPH [b], Lluïsa CURCOLL [b], Dolors SOLER [b], Laura PLA [b], José María TORMOS [b]
[a] *Knowledge Engineering and Machine Learning Group, Department of Statistics and Operations Research, Universitat Politecnica de Catalunya, Barcelona, Spain*
[b] *Institut Guttmann, Hospital de Neurorehabilitació, Badalona, Spain*

Abstract. In this paper, an integral Knowledge Discovery Methodology, named Clustering based on rules by States, which incorporates artificial intelligence (AI) and statistical methods as well as interpretation-oriented tools, is used for extracting knowledge patterns about the evolution over time of the Quality of Life (QoL) of patients with Spinal Cord Injury. The methodology incorporates the interaction with experts as a crucial element with the clustering methodology to guarantee usefulness of the results. Four typical patterns are discovered by taking into account prior expert knowledge. Several hypotheses are elaborated about the reasons for psychological distress or decreases in QoL of patients over time. The knowledge discovery from data (KDD) approach turns out, once again, to be a suitable formal framework for handling multidimensional complexity of the health domains.

Keywords. knowledge discovery from data, decision support and knowledge management, clustering, spinal cord injury, quality of life

1. Introduction

Spinal Cord Injury (SCI) is a modern epidemic [1]. Most of SCI cases have a traumatic origin and prevails in young people. Main causes are traffic, occupational and sporting accidents. There are about 1,000 new cases of traumatic SCI per year in Spain; between 140 and 160 in Catalonia. The prevalence of SCI is about 500 persons per million [2]. Considerable progress in the understanding of the pathogenesis and improvements in early recognition and treatments have occurred since Dr. Guttmann times. Currently, the SCI patients have a life expectancy similar to that of the general population, but disability persists throughout the patient's life. Maintenance or improvement of Quality of Life (QoL) is becoming a main goal, even for the WHO. Furthermore, the scientific community is now accepting that QoL can be formalized as a multidimensional construct depending on different objective as well as subjective aspects [3]. Indeed, in this work it was found that similar physical impairments can be followed by different psychological responses. Thus, it is important to understand the reasons and factors of a

[1] Corresponding Author: Ed. C5. Pta 2. Campus Nord, UPC, C. Jordi Girona 1-3, Barcelona 08034; E-mail: karina.gibert@upc.edu.

negative (or even depressive) psycho-emotional response as well as reasons of decreasing the QoL, to properly prevent them and give better assistance to patients.

Health-care environments are generally perceived [4] as rich in data. However, the performance of classical data analysis techniques with such complex domains use to be poor, and it is hard to get class-I scientific evidence. Knowledge Discovery from Data (KDD) was introduced in 1989 by Fayyad as *the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data* [5] under a global, iterative and multidisciplinary approach where crucial points are: interaction with the expert, data pre-processing, prior expert knowledge acquisition, Data Mining with a variety of non-classical techniques and post-processing of results (interpretation assessment) for producing *explicit* discovered knowledge. This allows a broader scope of much more complex problems to be analyzed than is possible under a more classical approach. In the study, a KDD methodology is used to extract knowledge about the most typical patterns of QoL over time using a sample of SCI patients. *Clustering based on rules* (ClBR) [6, 7] was successfully used in [8] to identify typical perceived QoL in a given moment. In this study, *ClBR by States* (ClBRxE) and new interpretation-assessment tools (for improving knowledge production) are used to discover trends of QoL over time. Some clinical hypotheses about QoL on SCI were elaborated and the theoretical doctrine corpus about psycho-social response patterns in SCI was reinforced.

2. Materials and Methods

2.1. PIE and Evaluation of QoL

All patients receiving care at the Institut Guttmann are followed-up after their clinical discharge. All of them are periodically re-assessed every 12–18 months by the Periodic Integral Evaluation (PIE). The PIE is conducted to permit early detection of some pathologies which, because of a baseline neurological lesion, might be asymptomatic till more advanced phases. Early detection can decrease complications, preventing long hospitalization, or even survival risks. The PIE also aims to improve autonomy levels, QoL and the social inclusion of the patients as much as possible. Medical, functional, psycho/neuropsychological and social aspects as well as health-education and health risk prevention are evaluated in the PIE with a total of 147 items. Institut Guttmann currently measures QoL by using several assessment scales, all included in the PIE, according to the multidimensional and multidisciplinar model of Schalock and Verdugo [3]. In the present study, a joint analysis by experts and knowledge engineers led to identify a subset of 32 items from the PIE dataset relevant for the QoL model of Schalock and Verdugo: **a) Emotional wellness**, measured through the instrument IBP (6 subscales); **b) Functional Autonomy**: Using the scores on the corresponding 19 items of the ICF, Disability and Health; including daily living activities (DLA), transfer, cognitive and social activities; **c) Social Inclusion** by means of the Social Scale of the Institut Guttmann (ESIG, 7 items).

2.2. Data Analysis Methodology

The target sample includes 109 patients with SCI followed up at the Institut Guttmann with at least three consecutive PIE between 2002 and 2008. First, basic descriptive statistics were obtained for all variables. Basic statistical methods were used to get preliminary information: histograms or bar charts to display variability, plots and

multiple box-plots to observe the relationships between pairs of variables, etc. Next, data cleaning was performed, including missing data treatment or detection of outliers; the quality of final results depends directly on this step. Decisions were taken on the basis of descriptive statistics and background expert's knowledge: Some redundant variables were eliminated and the final set of variables to be used was determined.

Then *ClBRxE* was applied to identify the more typical patterns of QoL over time in SCI. *ClBRxE*, implemented in the software KLASS [7, 9], generalizes ClBR [6, 7] to dynamical processes (both methods hybridate inductive learning (AI) and clustering (Statistics) for KDD in complex domains, using a Knowledge Base (KB) as a semantic bias for clustering): **i)** *ClBRxE* starts with local applications of ClBR (Ward's criteria, mixed Gibert's metrics [10]) to every PIE using KB, and maximising the ratio of inertias [11] to find the final number of classes; **ii)** Post-processing was based on *Traffic lights panel* (TLP) [12], very close to the expert knowledge: it is a symbolic abstraction of Class Panel Graph (CPG) [10] which compacts conditional distributions of variables through the classes. **iii)** Significance tests assessed relevance of variables versus classes (ANOVA, Kruskall-Wallis or χ^2 independence test). **iv)** These elements provided better interpretation-support to the expert in *conceptualization phase*, where they recognized the underlying profile of every class and labeled it with a semantic domain-concept**. v)** Finally, the knowledge discovered in every PIE is integrated into a global trajectories model [13]: trajectories of the patients are calculated as sequence of classes per State; frequencies of every trajectory are used to determine most probable trajectories. Those are graphed in the *trajectories diagram.* A small set of general and typical patterns of QoL over time is identified [7, 13].

3. Results

One hundred and nine persons with Spinal Cord Injury (89 men, 20 women, 76 paraplegic -42 with complete lesion-, 33 tetraplegic -18 with complete lesion) were considered. The mean age at time of injury was 44.52 years (SD=14.243 years). First, experts provided the prior KB. It described typical extreme cases in 12 rules as:

r1: if *patient is highly impaired and depressed with long time from injury* ***then*** *high dependency*
r2: if *high optimism and recent lesions and can perform DLA* ***then*** *independent for DLA*

ClBRxE was applied for the first three PIEs of every patient using KB and 4, 3 and 4 classes were identified for respectively the 1st, 2nd and 3rd PIE. In the conceptualization phase, experts analyzed the TLPs and the basic statistics per class of significant variables and could label all the classes. Then the trajectories diagram was built (Figure 1): the columns show the classes of every PIE with the labels provided by the expert. Green (at the top) is for wellness and independence; red (at the bottom), for high impairment and distress. Arrows display the trajectories of patients, and thickness reflects probability. Out of a total of 48 (4x3x4) possible trajectories, only 25 occurred. 52.3% of the sample follow one of the 4 most frequent trajectories ($p>0.05$):

T7 = (PIE1-Class55, PIE2-C63, PIE3-C59) ($p = 0.2477$): Persons with functional autonomy and psychological wellness maintaining this situation over the target period; many lived in couple.

T6 = (PIE1-C54, PIE2-C64, PIE3-C52) ($p = 0.1468$): Persons with high functional dependence who present heterogeneous psychological states at the beginning of the study. After one year they stabilize to a conformation state showing moderate distress, maintain good self-control, with no signs of anxiety or depression. They are tetraplegic patients with complete lesions.

T12 = (PIE1-C49, PIE2-C62, PIE3-C57) ($p = 0.0734$): Persons with functional autonomy and moderate distress, who persist in this situation. Most of them are young with a longer time from injury.

T4 = (PIE1-C49, PIE2-C62, PIE3-C56) (p = 0.055): Persons with functional autonomy and moderate distress at the beginning, who lose part of the functionality over time, and may also become distressed. Tended to be old persons, with longer time from injury.

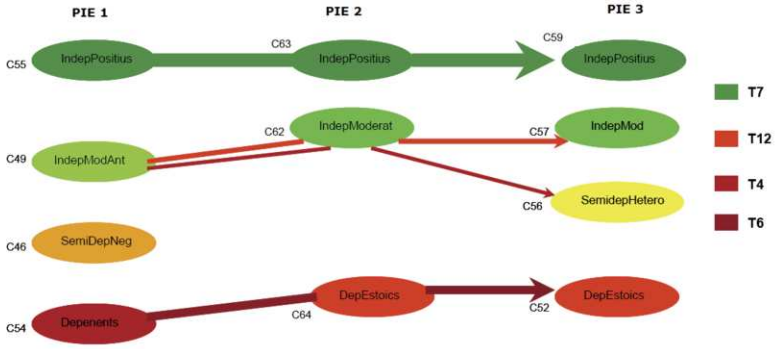

Figure 1. Trajectories diagram for the 3 PIEs (gamma = 0.05)

4. Discussion, Conclusions and Future Work

ClBRxE detects different groups of patients using their QoL perception over time. QoL in SCI is a very complex phenomenon, involving multidimensional factors as well as the patient's own perception, and successful analysis requires not only data, but also as much prior expert knowledge as possible. *ClBRxE* allows *partial* domain descriptions (in this case, just 12 rules involving only 7 variables were used). In this sense, it outperforms classical AI methods which require complete prior domain-descriptions and this is almost unreachable in complex domains because of the knowledge implicitly managed in expert's reasoning. *ClBRxE* guarantees interpretability as it permits integration of clinical knowledge into the analysis (KB), even if it is partial, and produces final patterns consistent with the prior expert knowledge. According to our opinion that *hybrid* techniques combining AI and statistics are better for *KDD* than pure ones, *ClBRxE* also outperforms, as all ClBR-like methods, classical mechanistic models in very complex domains, where most of the classical technical assumptions do not hold [6,7,10–15]. Several tools have been developed to assist the expert and properly insert him as part of the methodology itself. His participation, providing prior knowledge and interpreting results, is crucial to properly establishing the final patterns from a conceptual point of view. Finally, knowledge is expressed in terms of a subset of relevant variables, eventually useful for dimensionality reduction purposes. Figure 1 shows the diagram with the most frequent trajectories and it has proved to be a very intuitive tool for interpretation-support [13].

In the step **(i)** of *ClBRxE*, classes were identified by the different PIEs. They are characterized by the interaction between functionality and psychological status of the patient, whereas socio-demographic variables seem not to be relevant, except for time from injury, academic degree or living in couple. Although it is known [16] that maintaining high QoL is more difficult for the patients with higher impairment, psycho-emotional response seems not to be directly related with severity of the lesion as groups with moderate impairment and more distress were found (PIE1-C46, PIE3-C56).

More than half of the sample follows one of the four most frequent patterns of QoL over time recommended by the system, while the remaining 47.7% were scattered among

the other 21 observed patterns. The more typical trajectories were analyzed, although focus on an exception's analysis for understanding rare patterns is also possible. The four discovered patterns correspond well with professionally meaningful profiles. The general observed trend is stabilization over time. The more functionally dependent persons adopt a stoical attitude and tend to adjust to the disability and stabilize their psychological response. 78.8% of patients maintain (38.5%) or improve (43.2%) QoL; only 32% are old persons. Conversely, only 20 patients (old persons with longer time from injury) worsened; 70% of them are in T4. *ClBRxE* allowed to isolate age and time from injury as factors related with a decrease of QoL. *ClBRxE* seems an appropriate formal framework to discover useful knowledge from dynamical processes (which can be split into discrete steps) in very complex domains and produces much richer results compared with other approaches like [17] based on multiple comparisons and without interpretation-support tools, or [18] that merges all the observations over time into a single clustering thus forcing induction of a common set of classes for every PIE. The study also allowed to formulate various hypotheses about the improvement of perceived QoL and the reasons for distress, considering that coping can also depend on other factors like personality, social support perception, socio-familiar resources, dysfunctional couple [19], incomplete or medical lesions [19], pain. Next step is to record data required to confirm the generated hypotheses.

Acknowledgements. Subsecretaría de Estado de Servicios Sociales, Familia y Discapacidad del Ministerio de Trabajo y Asuntos Sociales; the Instituto de Salud Carlos III and the Agència d'Avaluació de Tecnologies i Recerca Mèdiques de la Generalitat de Catalunya.

References

[1] Kaur, H. et al (2006) Empirical study on applications of data mining techniques in healthcare. *Journal of Computer Science* 2(2):194–200.
[2] Institut Guttmann (2008) http://www.guttmann.com/index.aspx?opcio3=113&opcio2=11&opcio1=1.
[3] Schalock, R.L., Braddock, D., Verdugo, M.A. (2002) *Handbook on Quality of Life for Human Services Practitioners.* AAMR, Washington, DC.
[4] Kumar, V., Kumar, D., Singh, R.K. (2008) Outlier mining in medical databases. An Application of data mining in health care management to detect abnormal values presented in medical databases. *International Journal of Computer Science and Network Security* 8(8):272–277.
[5] Fayyad, U. (1996) From Data Mining to Knowledge Discovery: An Overview. AAAI Press, Menlo Park.
[6] Gibert, K. et al. (2005) KDD on functional disabilities. *Studies in Health Technology and Informatics* 16:163–168.
[7] Gibert, K. et al. (1998) ClBR and KDD in ISD. *Computación y sistemas* 1(4):213–227.
[8] Curcoll, M.L. et al. (2007) Deliverable 2. Project Laboratorio de Medidas Potenciadoras de la Autonomía, Satisfacción Personal y Calidad deVida de las personas con LM o Daño Cerebral Adquirido.
[9] Gibert, K., Nonell, R. (2008) Pre and post-processing in KLASS. In *Proceedings of the DM-TES'08 Workshop at iEMSs 2008*, vol. III, 1965–1966.
[10] Gibert, K. et al. (2005) KDD with clustering: impact of metrics. *Neural Networks World* 15(4):319–326.
[11] Gibert, K. et al. (2008) Data mining for environmental systems. In *IDEA series*, v3, Elsevier, 205–228.
[12] Gibert, K. et al. (2009) The role of KDD support-interpretation tools in the conceptualization of medical profiles: An application to neurorehabilitation. *Acta Informatica Medica* 8(2):170–180.
[13] Gibert, K. et al. (2009) KDD in a WWTP with ClBRxE. *Environmental Modelling and Software* (in press)
[14] Gibert, K. et al. (2008) Response to TBI-neurorehabilitation through an AI&Stats hybrid KDD methodology. *Medical Archives* 62(3):132–135.
[15] Gibert, K., Sonicki, Z. (1999) Classification based on rules and medical research. *Journal of Applied Mathematics and Stochastic Analysis* 15(3):319–324.
[16] Elfström et al. (2002) Effects of coping on psychological outcome when controlling for background variables: A study of traumatically spinal cord lesioned persons. *Spinal Cord* 40:404–415.
[17] Yilmaz et al. (2005) Long-term follow-up of patients with spinal cord injury. *Neurorehabilitation and Neural Repair* 19:332–337.
[18] Sugar, C.A. et al. (1998) Empirically defined health states for depression from the SF-12. *Health Services Research* 33:911–928.
[19] Curcoll, M.L. (1999) Lesiones Medulares y Rehabilitación. *Mapfre Medicina*.