The Practice of Patient Centered Care: Empowering and Engaging Patients in the Digital Era
R. Engelbrecht et al. (Eds.)
© 2017 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/978-1-61499-824-2-38

DisEpi: Compact Visualization as a Tool for Applied Epidemiological Research

Arriel BENIS [a,1] and Moshe HOSHEN [a]
[a] *Clalit Research Institute, Chief Physician's Office, Clalit Health Services, Tel-Aviv, Israel*

Abstract. Outcomes research and evidence-based medical practice is being positively impacted by proliferation of healthcare databases. Modern epidemiologic studies require complex data comprehension. A new tool, DisEpi, facilitates visual exploration of epidemiological data supporting Public Health Knowledge Discovery. It provides domain-experts a compact visualization of information at the population level. In this study, DisEpi is applied to Attention-Deficit/Hyperactivity Disorder (ADHD) patients within Clalit Health Services, analyzing the socio-demographic and ADHD filled prescription data between 2006 and 2016 of 1,605,800 children aged 6 to 17 years. DisEpi's goals facilitate the identification of (1) Links between attributes and/or events, (2) Changes in these relationships over time, and (3) Clusters of population attributes for similar trends. DisEpi combines hierarchical clustering graphics and a heatmap where color shades reflect disease time-trends. In the ADHD context, DisEpi allowed the domain-expert to visually analyze a snapshot summary of data mining results. Accordingly, the domain-expert was able to efficiently identify that: (1) Relatively younger children and particularly youngest children in class are treated more often, (2) Medication incidence increased between 2006 and 2011 but then stabilized, and (3) Progression rates of medication incidence is different for each of the 3 main discovered clusters (aka: profiles) of treated children. DisEpi delivered results similar to those previously published which used classical statistical approaches. DisEpi requires minimal preparation and fewer iterations, generating results in a user-friendly format for the domain-expert. DisEpi will be wrapped as a package containing the end-to-end discovery process. Optionally, it may provide automated annotation using calendar events (such as policy changes or media interests), which can improve discovery efficiency, interpretation, and policy implementation.

Keywords. Visual Data Mining, Clustering, Heatmap, Epidemiology, Public Health Informatics, Facilitation, ADHD

1. Introduction

During the last decades, data collection and storage have been facilitated by a large number of Computer Science and Computer Engineering evolutions. In the Healthcare realm, the numbers of datasets and their diversity have grown at a dramatic rate. These changes facilitate rapid progress in outcomes research and evidence-based medical practice. The Knowledge Discovery in Databases (KDD) [1] framework provides mathematical and Artificial Intelligence based approaches, such as Data Mining and particularly, Machine Learning and Information Visualization [2], for extraction of

[1] Corresponding author, Dr. Arriel Benis, Clalit Research Institute, Chief Physician's Office, Clalit Health Services, Arlozorov 101, Tel-Aviv 6209804, Israel; E-mail: arriel.benis@gmail.com.

useful and non-trivial information about patterns and relationships existing within the huge available data that might otherwise be missed by classical statistical approaches.

Public Health Informatics (PHI) is a link between Public Health (PH) as a discipline of caring for the community health in general, risk groups in particular, and Informatics, which is a discipline residing in the combination of Computer and Information Sciences and Engineering. PHI focuses on development of methodologies and applications of Informatics systems in PH and particularly in Epidemiology and its application subfields (surveillance, prevention, preparedness, and health promotion).

One of the main roles of PH Informaticists is to support KDD through tasks, such as indicating behavioral patterns in administrative and bio-clinical data. A large number of statistical techniques and Machine Learning methods can be used to fulfill these objectives. One main use of the data mining step of KDD in PH is supplying Management-level decision makers with simple tools to support the development of new healthcare policies. Data and Information Visualization allows the domain expert (e.g. the biologist, the physician, the public health expert) to comprehend, better and faster, numerical and textual data, based on human innate abilities and basic learning. Summarized and synthesized information lend themselves to rapid digesting of data and making relevant decisions. Viewing visualizations saves time and improves understanding. These representations and interactions allow easy efficient acquisition of complex information. Enriched data may also be readily placed online with appropriate visualizations for open use [3, 4].

In the present paper we introduce a novel method, DisEpi *(Discovery in Epidemiology)*, for visually exploring epidemiological data in order to detect trends within sub-populations, so as to allow focused interventions.

DisEpi is an efficient Information Visualization tool. The tool receives, as input, a large administrative dataset (e.g. individual level demographics, socio-economic, medical, and therapeutic data) for a large number of individuals, and, without explicit modelling, creates an easy, compact visualization [5] of the data at the population level affording the executive professional rapid discovery of (1) Links and associations between events, (2) Association changes over time, and (3) Clusters of characteristics with similar trends. Thus, "DisEpi" is designed to support PH practitioners and researchers by providing semi-automated hypothesis generation and by reducing time to investigation and interpretation.

2. Background

Incorporation of very large databases in epidemiological studies involves complex integration of demographics, socio-economic, medical, and pharmacological data collected over the course of several years. Thus, epidemiological data mining is designed to detect and describe patterns, trends, and relations in medical data, allowing definition of specific research questions, as part of the knowledge discovery process. Various research studies have been conducted to support classical descriptive and predictive epidemiological studies [6, 7, 8, 9, 10], each one with a specific point-of-view.

DisEpi applies Abstraction, Reformulation, and Approximation [11] concepts and techniques which provides the epidemiologist and PH professional with compact visual information at the population level. DisEpi thus overcomes some limitations of the classical statistical approaches. It utilizes Hierarchical Clustering (HC), a well-known

unsupervised machine learning approach comprising of successive agglomeration of cases with minimal distances between them as defined by suitable metrics [12]. DisEpi uses heatmaps [13], a symbolic gradient reflecting the epidemiological metrics' values with HC as one dimension (e.g. columns) and time or time-associated variables as a second dimension (e.g. rows) and where color shades as the third dimension reflect disease levels.

3. Material and Method

3.1. Material

DisEpi was applied in the context of a study of Attention-Deficit/Hyperactivity Disorder (ADHD). ADHD is one of the most commonly diagnosed mental conditions in children. It may have lasting social, psychological, educational and clinical effects throughout life [14].

The data were extracted from electronic health records (EHR) of the Clalit Health Services (Clalit), which is the largest healthcare payer/provider organization in Israel, covering more than 50% of the population, over ,4400,000 members. The study population focused on 1,605,800 children aged between 6 and 17 years, from 2006 and 2016.

The children's estimated relative ages in year in school (class) were split into thirds of the calendar year, with the youngest third born between August and November, the middle third born between April and July, and the oldest third born between December and March. The study was approved by the Clalit Health Services Ethics Committee [15].

3.2. Method

The output of DisEpi is a compact Visualization combining the graphical representation of the automatically generated HC and a heatmap, wherein a color gradient reflects either (a) incidence or (b) prevalence values over the sub-time ranges over the research time-range.

As a first step of the knowledge discovery process, we built an abstraction process discretizing the available data of each attribute into several classes, e.g.: age groups in two year spans, number of siblings (0-1, 2-3, and 4 or more), and ordinal month of birth relative to the school-year (categorized into youngest, middle, and oldest thirds).

As a second step, we approximated all attributes, from individual child data to sub-population snapshot values, describing respective statuses. These approximations were computed by using epidemiological metrics such as incidence and prevalence, at the relative age thirds' level.

As a third step, we computed a Euclidean distance matrix between all epidemiological metrics values describing the classes of each attribute (one value per attribute per relative age in class per year) within which we applied unsupervised HC. .

The fourth step of DisEpi discovery flow is the reformulation process. This process is based on the use of a heatmap for each epidemiological metric (e.g. incidence and prevalence).

We used R version 3.3.1 [16] for the statistical analysis. R Package doParallel [17] was used to maximize computing efficiency given the large data size. R Package gplots [18] was used to perform automatically the HC and then to draw the heatmap.

4. Results

In the context of ADHD, DisEpi allows the domain-expert to visually analyze data mining results in a summary snapshot (**Figure 1**). Looking at the snapshot of ADHD medication usage, he/she is able to grasp the following results, simultaneously: (1) Relatively younger children in class are treated more than older; (2) First medication incidence increased between 2006 and 2011 but then stabilized; (3) Progression rates of medication incidence are different for each of 3 principal clusters.

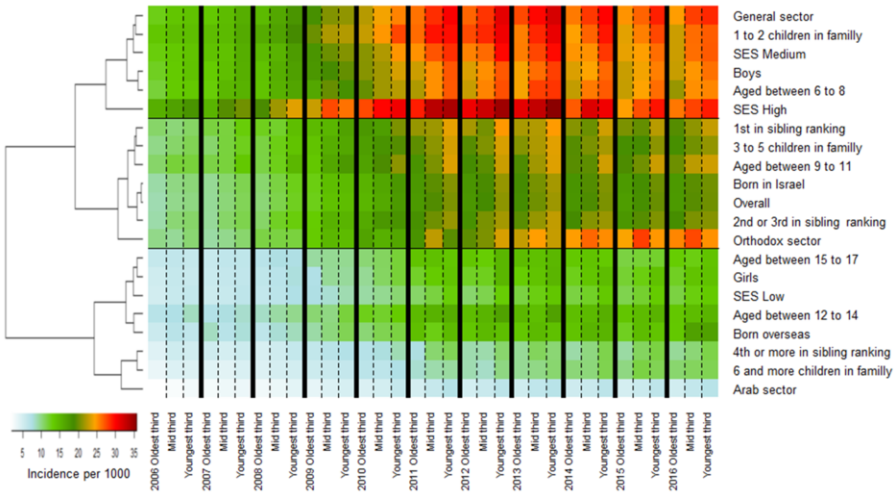

Figure 1. DisEpi Visualization for ADHD medication incidence over the explored attributes.

The pharmaco-epidemiological goal of this DisEpi experimentation was to identify characteristics and medication treatment patterns of Clalit's youngest population treated for ADHD and to compare these patterns to sociological and demographical data, such as relative age in class, sex, ethnicity, family size, sibling order, and other socioeconomic status attributes. DisEpi results are similar to those discovered in a previous research using classic statistical approaches [15], but are delivered in a more rapid, concise, and user-friendly way.

Furthermore, in the previous research on the same data [15], by using *a priori* based data analysis, one could not readily identify the similarity of the children population born overseas to the lately treated group. This means that DisEpi supports new epidemiological discoveries which must be later validated by the domain-expert.

5. Conclusions and perspectives

In this paper, we presented DisEpi, a tool allowing for the identification of clusters of population characteristics having similar trends, and thus defining profiles of patients,

with relatively short delays (a few days instead of a few months). It allows users to deal with very large medical numerico-symbolic data by reducing their dimensionality and complexity, and making them more accessible to PH decision makers. Therefore, it can support development of health policies by reducing delay by providing the medical decision makers a "data/analysis" integrated tool.

DisEpi will be wrapped as a package containing the end-to-end discovery process and will provide automated annotation using calendar events (e.g. "news"), which will reduce both time to discovery, interpretation and policy implementation [19].

References

[1] U. Fayyad, . G. Piatetsky-Shapiro, P. Smyth and T. Widener, "The KDD Process for Extracting Useful Knowledge from Volumes of Data," Communications of the ACM, vol. 39, pp. 27--34, 1996.

[2] D. Keim, "Information Visualization and Visual Data Mining.," IEEE Transactions on Visualization and Computer Graphics 2002, vol. 8, no. 1, pp. 1-8, 2002.

[3] J. Bertin, Semiology of graphics, Madison, Wisconsin: University of Wisconsin Press, 1983.

[4] E. R. Tufte, The Visual Display of Quantitative Information, 2nd ed., Cheshire: Graphics Press, 2001.

[5] B. Shneiderman, "The eyes have it: A task by data type," in Symposium Visual Language, 1996.

[6] J. Quentin-Trautvetter, P. Devos, A. Duhamel, R. Beuscart and Qualidiab Group, "Assessing association rules and decision trees on analysis of diabetes data from the DiabCare program in France," Stud Health Technol Inform, no. 90, pp. 557-61., 2002.

[7] P. Valle, O. Flaten, G. Lien, M. Koesling, C. Carrol and M. Ebbesvik, "Data Mining, a useful tool in veterinary epidemiology?," in Proceedings of the 10th International Symposium on Veterinary Epidemiology and Economics, Vina del Mar, Chile, 2003.

[8] C. R. Twardy, A. E. Nicholson, K. B. Korb and J. McNeil, "Data Mining Cardiovascular Bayesian Networks," Melbourne, 2004.

[9] A. Benis and M. Courtine, "Biomarkers discovery in medical genomics data," in Advances in experimental medicine and biology, vol. 696, Springer, 2011, pp. 327-34.

[10] S. N. Rajan, A. K. Sinha and J. B. Singh, "The Study of Knowledge Discovery with Spatial Data Mining in Epidemiology Database," International Journal of Engineering Research & Technology, vol. 1, no. 6, August 2014.

[11] L. Saitta and J.-D. Zucker, Abstraction in Artificial Intelligence and Complex Systems, 1 ed., New York: Springer-Verlag, 2013, p. 484.

[12] L. Kaufman and P. Roussew, Finding Groups in Data - An Introduction to Cluster Analysis, Wiley-Science Publication John Wiley & Sons., 1990.

[13] L. Wilkinson and M. Friendly, "The History of the Cluster Heat Map," vol. 63, no. 2, pp. 179-184, 2009.

[14] J. Biederman, "Attention-deficit/hyperactivity disorder: a selective overview," Biological psychiatry, vol. 57, no. 11, pp. 1215-20, 2005.

[15] M. Hoshen, A. Benis, K. M. Keyes and H. Zoëga, "Stimulant use for ADHD and relative age in class among children in Israel," Pharmacoepidemiology and Drug Safety, 2016.

[16] R Core Team, R: a language and environment for statistical computing, 2017.

[17] R. Calaway , Revolution Analytics , S. Weston and D. Tenenbaum, doParallel: Foreach Parallel Adaptor for the 'parallel' Package, 2015.

[18] G. R. Warnes, B. Bolker, L. Bonebakker, W. H. A. Liaw, W. Huber, A. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz and B. Venables, gplots: Various R Programming Tools for Plotting Data, 2015.

[19] K. J. S. Lim and S. Douglas, "Feeling the Market's Pulse with Google Trend," 28 September 2014.