Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230119

Greek Hospital Data Mining and Analysis

Maria CHINTIROGLOU[a,1], Haralampos KARANIKAS[a] and Sotiris TASOULIS [a]
[a] *Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece*

Abstract. Monitoring the performance of hospitals is a crucial issue related both with the quality of healthcare services and with country's economy. An easy and trustful way of evaluating health systems is through key performance indicators (KPIs). Such indicators are widely used for the identification of gaps in the quality or efficiency of the services provided. The main aim of this study is the analysis of the financial and operational indicators at hospitals in the 3rd and 5th Healthcare Regions of Greece. In addition, through cluster analysis and data visualization we attempt to uncover hidden patterns that may lie within our data. The results of the study support the need for re-evaluation of the assessment methodology of Greek hospitals to identify the weaknesses in the system, while evidently unsupervised learning exposes the potential of group-based decision making.

Keywords. Hospital Data, Data Analysis, eHealth, key performance indicators.

1. Introduction

Over the years, various methods have been sought to ensure the proper performance of hospitals and increase their efficiency in Greece. Despite the scientific and technological development, the health systems of many countries still present weaknesses. It is estimated that 40%-80% of a country's total expenses are allocated to hospitals [1]. Therefore, it was deemed necessary to adopt policies for evaluating hospitals' operation.

In Greece, one of the methods for monitoring the operation of hospitals is by certain established KPIs. These indicators can be divided in two main categories: financial and operational [2]. The financial indicators include annual expenditure for raw and auxiliary materials (pharmaceuticals, hygiene supplies, orthopedic equipment, reagents, etc.), consumables (gas, fuel, etc.) and detailed data on salaries, payments, and revenues for every hospital. The hospital activity (operational) data include the number of inpatients and the total hospitalization days, the number of outpatients and the emergency services for every hospital. In this study, data from hospitals in the 3rd and 5th healthcare regions of Greece were used. According to the data of 2014, the corresponding indicators have been produced and then evaluated. In addition, particular indicators have been combined to construct a new data matrix, upon which unsupervised learning have been applied to discover groups of hospital with similar respective behavior. While the results are specific, this method is rather generic and can also be employed in other countries, setups, or data.

[1] Corresponding Author: Maria Chintiroglou, Department of Computer Science and Biomedical Informatics, University of Thessaly, Papasiopoulou 2-4, Postal Code: 35131, Lamia, Greece; E-mail: mhintirog@gmail.com

2. Methods

The provided data, financial data of the hospitals (1), data corresponding to the number of beds (2), data relative to the inpatient number (3), and data concerning the outpatients (4), underwent a pre-screening process, to facilitate the analysis of the indicators and the clustering procedure.

Our basic workflow was divided in two main parts:

- The production and the evaluation of indicators.
- The combination of indicators, and the subsequent cluster analysis.

This study estimated and examined the following indicatively KPIs: average cost per patient, defined as the total expenditure (expenditure for raw and auxiliary materials and consumables etc., excluding payroll) of a hospital, divided by the number of hospitalized patients; average cost per hospital day, defined as the total expenditure (expenditure for raw and auxiliary materials and consumables etc., excluding payroll) of a hospital, divided by the total number of hospital days; average drug cost per patient, defined as the total pharmaceutical expenditure of the hospital, divided by the number of hospitalized patients. Initially, we calculated the total number of patients, days of hospitalization, and of beds, needed to produce the corresponding indicators. Linear regression analysis was performed to select between the number of patients and the number of days of hospitalization as the most suitable for the calculation of each indicator. The most appropriate indicators are being utilized to generate the dataset for the unsupervised analysis. The resulting dataset is constituted by the 7 calculated indicators along with a variable with the total number of beds per hospital. It is important to mention that for the creation of the dataset we only considered general hospitals to focus our analysis on the dominant type of hospitals found at the data at hand. The inclusion of other types could provide misleading results due to the major fundamental differences between hospital types. To uncover hidden patterns that may lie within our data we utilize clustering. Initially, we separate the data into group of hospitals that share similar characteristics and then investigate the critical factors that determine the retrieved groups (clusters). Cluster analysis, combined with visualization of the results, is a very popular and well-established methodology usually used to make sense of large data volumes, finding wide applicability even on recent applications in Bioinformatics and Big Data [3, 4, 5]. Although there are several recent advancements in clustering algorithms [6, 7], for the task at hand we utilize the traditional “k-means” due to its simplicity and control over its parameters. An adequate amount of data would allow the utilization of more sophisticated techniques that promote explainability even further. To this end, the available dataset’s manageable scale allows multiple algorithmic executions that can lead to relative optimal results, without the need for advanced variations intended for complexity reduction and appropriate initialization.

3. Results

In this section, the charts of the most important and most discussed hospital indicators, namely average total cost per patient and pharmaceutical cost per patient (Fig.1 and Fig.2, respectively), are indicatively presented.

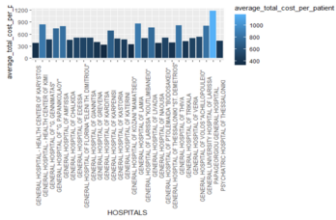

Figure 1. Average total cost per patient.

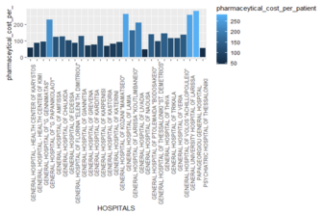

Figure 2. Pharmaceutical cost per patient.

From Figure 1 it is clear that Papageorgiou Hospital has the highest average total cost per patient and, on the opposite side, Thiva Hospital has the lowest. In general, most of the hospitals have similar average inpatient costs except for some that deviate highly. It is noteworthy that Kymi General Hospital has a fairly high average cost, due to its restricted facilities and more specifically based on its low number of beds. This could be possibly attributed to the aeromedical evacuation that may be required, as Kymi is a relatively remote place in Greece. From Figure 2 it can be observed that the largest number of hospitals show similar average costs for drugs with only five hospitals differing in a larger value. Papageorgiou General Hospital shows the highest pharmaceutical cost per patient, and opposingly, Naoussa General Hospital shows the lowest. Pharmaceutical costs are certainly influenced by many factors such as the type of cases that each hospital deals with, which may explain the large variations between hospitals. Moreover, the average length of stay is around 4 days in all hospitals apart from the Psychiatric Hospital. This is reasonable because psychiatric cases probably need a greater length of stay and care [8].

The occupancy of the beds, as it was observed, is greatly increased in Karystos and Kymis Hospitals. After all, they are the hospitals with the lowest number of beds. In the hospitals with the largest number of beds, such as Papageorgiou Hospital and Papanikolaou Hospital, the bed occupancy is much lower. Additionally, very low bed occupancy is observed in the University Hospital and the Psychiatric Hospital.

To this end we employ the dataset described in “Methods” Section. Initially, we perform range normalization across all variables bounding their values in [0,1], while to determine the number of clusters parameter we utilize the elbow method. In figure 3 we see a line chart of the SSE for each value of k. Based on the results and on the fact that we need to retain a relatively low number of clusters for straightforward interpret-ability we set k = 4. To visually investigate the uncovered data structure we perform Principal Components Analysis (PCA) to project the data onto the first two principal components. As shown, in Figure 3 we can clearly distinguish identified clusters.

Figure 3. Principal Components Analysis (PCA).

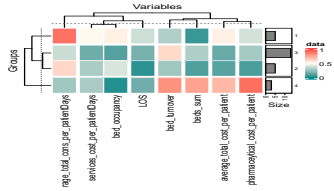

Figure 4. Average values for each normalized variable per cluster.

To identify the aspects of each retrieved cluster we present the corresponding average values of each variable in Figure 4, allowing us to discriminate major differences along clusters. To enhance the visualization both retrieved clusters and variables are grouped through agglomerative clustering to re-positioning them according to their in-between similarities. The size of each cluster is also reported through a bar plot.

Finally, we attempt to relate the clustering result with the hospital geolocation in order identify the relation of the extracted hospital factors with their geographical location. Figure 5 illustrate the location of all hospitals colored according to the respective cluster label of Figure 3.

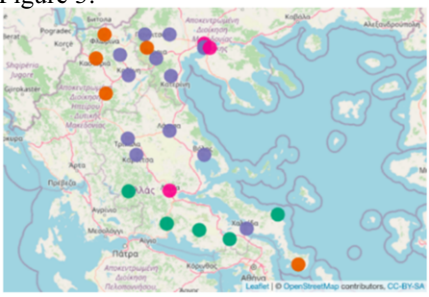

Figure 5. Hospitals Geographic Location.

From the above clustering we could proceed to draw some conclusions. First of all, we conclude that hospitals belonging to nearby cities have common characteristics in their mode of operation based on their indicators. This could be either because of the life of the citizens in each area or because of following a common strategy in the operation of the nearby hospitals. We also observe that the number of beds has a significant impact on the performance of hospitals. We observe that hospitals with the largest number of beds have lower costs for services and for materials and consumables and higher costs for drugs while hospitals with the smallest number of beds have higher costs in materials and consumables and lower costs for drugs. This could be due to the fact that medical cases requiring more serious care such as more serious operations or admissions to ICU etc. are transferred to the larger and more centralized hospitals in the country. This affects the larger hospitals that must manage more serious cases and thus increasing their overall costs per patient and their overall costs for drugs. Smaller hospitals on the other hand are burdened with costs for materials and consumables such as costs for patient transport etc. [9]. Furthermore, it seems that bed occupancy, a very important indicator for the performance of a hospital, is affected by the size of the number of beds, as hospitals with the largest number of beds do not have a problem with occupancy while hospitals with the lowest number of beds seem to have an issue with their bed occupancy which may also support our previous conclusion about the transfer of patients from smaller to larger hospitals also due to their bed occupancy.

4. Discussion and Conclusions

Hospitals have to face many cases of patients daily, such as outpatients, emergency cases, etc., with high monitoring needs. Due to the economic crisis, the workload requires faster and more trustworthy data processing. Thus, a way to monitor hospitals' data in a faster and more reliable fashion, is with the use of indicators. Furthermore, extracting knowledge from this valuable data is a major piece of research. The use of clustering

algorithms can contribute to optimal and faster knowledge discovery. The need for monitoring and evaluation of health units is, therefore, urgent. Indicators are a reliable way to achieve this. From the indicators calculated above, we can conclude that in general the larger hospitals face the most workload and have the highest costs apart from some exceptions such as Kymi General Hospital which is a small hospital but has high costs for services, materials, and consumables. At the same time, it can be concluded that the average length of stay in almost all hospitals is not very long, with an average of 3-5 days, except for Psychiatric hospitals, where patients stay on average 20 days.

Furthermore, it appears that hospitals that belongs in nearby geographical areas exhibit similar characteristics in their operation. Hospitals that have similar number of beds also show similar performance. Finally, we could conclude that smaller and relatively remote hospitals face more difficulties in their performance than hospitals located in larger and more central cities of the country. Finally, all the above observations and conclusions allow us to gain an insight into the quality of care in each hospital. The bed occupancy rate, for example, which was mentioned above, is a very important indicator not only for the functioning of a hospital, but also for the quality of care it provides. Hospitals with a high bed occupancy rate, above average and above permissible limits, cannot provide the same quality level of care to their patients as hospitals that do not face such a high workload.

References

[1] Rahimi H, Khammar-nia M, Kavosi Z, Eslahi M. Indicators of hospital performance evaluation: a systematic review. International Journal of Hospital Research. 2014 Dec;3(4):199-208.

[2] Christodoulakis A, Karanikas H, Billiris A, Thireos E, Pelekis N. “Big data” in health care Assessment of the performance of Greek NHS hospitals using key performance and clinical workload indicators. Journal: ARCHIVES OF HELLENIC MEDICINE. 2016;33(4):489-497.

[3] Ianni M, Masciari E, Mazzeo GM, Mezzanzanica M, Zaniolo C. Fast and effective big data exploration by clustering.Future Generation Computer Systems.2020;102:84-94.doi: https://doi.org/https://doi.org/10.1016/j.future.2019.07.077

[4] Kern M, Lex A, Gehlenborg N, Johnson CR. Interactive visual exploration and refinement of cluster assignments. BMC Bioinformatics. 2017 Sep;18(1). doi: https://doi.org/10.1186/s12859-017-1813-7

[5] Ostaszewski M, Kieffer E, Danoy G, Schneider R, Bouvry P. Clustering approaches for visual knowledge exploration in molecular interaction networks. BMC Bioinformatics. 2018 Aug;19(1).doi: https://doi.org/10.1186/s12859-018-2314-z.

[6] Sieranoja S, Fra¨nti P. Fast and general density peaks clustering. Pattern Recognition Letters. 2019 Dec; 128:551-558. doi: https://doi.org/10.1016/j.patrec.2019.10.019.

[7] Tasoulis S, Pavlidis NG, Roos T. Nonlinear dimensionality reduction for clustering. Pattern Recognition. 2020 Nov;107:107508. doi: https://doi.org/10.1016/j.patcog.2020.107508.

[8] Bressi SK, Marcus SC, Solomon PL. The impact of psychiatric comorbidity on general hospital length of stay. Psychiatric Quarterly. 2006 Sep;77(3):203-209.doi: https://doi.org/10.1007/s11126-006-9007-x

[9] Kontodimopoulos N, Nanos P, Niakas D. Balancing efficiency of health services and equity of access in remote areas in Greece. Health policy. 2006;76(1):49-57.