EFFICIENCY ANALYSIS IN HEALTH CARE BY MULTIVARIATE CLASSIFICATION METHODS

Francesco D. d'OVIDIO¹, Domenico LEOGRANDE², Rossana MANCARELLA³, Domenico VIOLA⁴

¹Department of Economics and Mathematics
francescodomenico.dovidio@uniba.it
²Department of Business and Private Law Studies
domenico.leogrande@uniba.it
³Department of Business and Private Law Studies
rossanamancarella@gmail.com,
University of Bari “Aldo Moro”, Largo Abbazia Santa Scolastica, 70124 Bari, Italy

Abstract
Due to the rapidly rising costs and to limited financial resources, in recent years, the European health institutions were increasingly encouraged to implement rational tools in order to better manage available resources and to make efficient choices in order to have the best cost-benefit ratio. Due to the strategic role of the Health Care sector is very important in an economic system, the fulfilment of the European standards government requires to develop tools able to fastly identify the inefficiency situations.

This contribution aims to study and investigate efficiency in health organization using multivariate methods of data mining (by example, segmentation analysis and neural networks).

Its practical interest will be directed in particular to the assessment of organizational appropriateness in health care, in order to evaluate the incidence of the “day hospital” and “day surgery” procedures, analysing their relevance in the health system as well as their appropriate level in the Apulian region.

One of the results of such analysis is to understand some decisional mechanism of the Health Care management, as well as the structure of inefficiency in the health network.

Key words; health system, segmentation analysis, cost benefit ratio.

Introduction: framework of the research
This paper aims to evaluate statistical techniques (such as segmentation analysis sample) the quality of hospital services, paying particular attention to a specific aspect: the organizational appropriateness, referring to the use of the instrument of the “day hospital”.

The organizational appropriateness of health care, particularly in hospital admissions, has recently become a topic of great interest not only in Italy. The overall intention is to move, where possible, assistance from the hospital to the area and avoid unnecessary hospitalizations, or they can be adequately addressed by other means of assistance. In light of this, the objective of this work is to verify, by means of statistical procedures, the organizational appropriateness of hospital admissions made in public and private institutions in a local area, from 2003 to 2008 (latest correct data available at this time, but the identified methodology can be easily applied also to the most recent data). This survey was carried out by using data provided by the Regional Health (ARES) of the Apulia region, consisting of all the information (except identity data) that can be extrapolated from hospital discharge records (HDR) of patients admitted in that region within the relevant time for the survey.

The data included in the HDR (now including more than 65,000 diagnostic codes) are grouped according to different classifications: the DRG system (Diagnosis related groups), and the grouping system ACC (Aggregated Clinical Codes, now known as
CCS\(^1\), based on clinical criteria of affinity and very useful for epidemiological analysis and health planning, but generally less appropriate DRG for the purpose of economic planning. These ACC are too many for this statistic experiment; then, in this context, we decided to conduct the survey on the 24 ACCs related to cardiovascular disease, with a number of admissions in public and private hospitals of all Apulian area amounting to 1,369,216 (anyway, a number higher than the patients actually hospitalized, due to multiple admissions).

**Hospital Admissions and Day Hospital**
The types of hospital admission, different than the ordinary mode, are: Day Hospital (distinguished between diagnostic, therapeutic and rehabilitative) and Day Surgery.

The day hospital is a particular type of admission, once typical of hospitals for acute care or with rehabilitation functions; it takes place only daytime, and it consists of one or more programmed “accesses”, all recorded in a single patient record (opened at the first admission). The day hospital may include: an integrated assessment of most specialists about the clinical situation of the patient; the provision of specialized services of health care, both medical and surgical, without the need of overnight; monitoring for side effects of particular drugs; the execution of diagnostic procedures, also invasive, for which a brief period of observation is indicated. Often the actions appropriately held during accesses to a day-care may be inappropriate when provided for inpatient (due to the least effort required), or in-patient treatment (because of the complexity of benefits). The key difference between the two admission schemes is the need for overnight in hospitalization; compared to outpatient treatment, instead, the day hospital has the prerogative of the time needed to carry out the performance and especially the need of the patient to be subjected, in this access time, to more medical services or nursing and/or to require a some hours of observation.

Finally, the day surgery is the scheme for the admission of patients who require surgery or invasive diagnostic procedures that, although resolved without overnight in hospitalization, require the use of the operating surgery room (its operations and procedures have to be included in the classification of procedures, in order to identify the surgical DRGs).

Due to reasons surely valid by a clinical standpoint, the ordinary hospitalization is the preferred choice by health care professionals, relying almost 86.5% of cases, while the Day Hospital (or Day Surgery) is selected only in the remaining 13.5% of cases. However, in the observed period the proportion of admissions in Day Hospital increased significantly, from less than 10% until 16%. Also the ordinary admissions increased by an amount roughly similar (19,263 versus 18,717 DH), but the growth trend is much larger in the last time series: set equal to 100 the respective number of admissions in 2003, in 2008 the inpatient admissions are around 110, while those in DH are more than 190.

**Characteristics of the hospitalizations, survey assumptions and sampling procedure**

3.1 - Characteristics of the hospitalizations
The hospitalization distribution by sex is almost even, but with a slightly greater weight for males; almost similar is also the choice of DH. Instead, the distribution of hospitalized by age group isn’t homogeneous, for obvious medical reasons, as the percentage of patients under 25 years is very low (about 2%), and grows in the range 25-54 (15%). About 45% of admissions are concentrated in the age group 55-74 years, while in the last age group (over 75 years) they are less than 38%: but this age group in the population is less large. Indeed, many of these diseases develop with age or as a result of diseases previously developed. The presence of young people, although a small percentage may be due to complications of childhood diseases or very rare congenital cardiac malformations.

As regards the district in which hospitals are located, the district of Bari focuses almost 38% of the total, followed once again by the districts of Foggia (19, 4%) and Lecce (17.6%). Just 5% of hospitalizations concerns the district BAT (Barletta-Andria-Trani), and slightly more (8.3%) is registered in the district of Brindisi. The hospitals of the Bari district are also those that frequently resort to DH (more than 21% of the admissions), while the lower proportion of hospitalizations in Day Hospital can be seen once again in the BAT. Instead, the hospitals of the province of Brindisi point out a fair recourse to the Day Hospital.

3.2 – Survey design and sampling procedure
In order to explore the organizational appropriateness of the health care procedures, identifying the determinants of lower rates of Day Hospital admissions respect to the overall rate, a classification analysis was designed in this survey.

\(^1\)The ACC (301 diagnostic ACCs and 231 ACCs of procedures, in ICD-9 CM) were developed since 1993 by the US Agency for Healthcare Quality and Research (AHQR), in order to group the ICD 9 CM in a restricted set of relatively homogeneous classes, useful for researchers and analysts. This classification was progressively obtained by modifying the previous CCHPR (Clinical Classifications for Health Policy Research), following clinical criteria of homogeneity and the frequency of hospital discharges. It was periodically updated following the the ICD. Since 1999 the system was computerized and renamed “Clinical Classifications Software”.
The underlying assumption behind this analysis is that, if the decision to provide a *inpatient* or an *outpatient* hospitalization only depends on clinical grounds (ie, the diagnoses), patients will be grouped homogeneously only under the same diagnosis (here synthesized by ACC). If, instead, other elements are involved in classification, different than the clinical choices (for example, a greater or lesser economic convenience, or unavailability of technical or human resources, or other), these elements are to be highlighted and their positive or negative influence on the organizational appropriateness can be investigated, removing the causes of negative discrepancies or acquiring the “good practices” that gave the better results.

In statistics, a large number of observations generally leads to optimal results: therefore an analysis conducted on the entire dataset of hospitalizations (over one million three hundred thousand cases) seems the most effective choice. But the strong asymmetry of the data distribution (by hospitalization type) gives a hard penalty to the classification algorithms\(^2\). This problem particularly the least represented categories: and indeed the discriminating power of the method was very low for the DH admissions.

In order to obtain better results than those possible with the original data, then, it is necessary to build a new set of data, as much as possible balanced, ie presenting roughly even number of inpatient hospitalizations and DH admission. Therefore, a new dataset was built by the random selection of a statistically representative sample with unequal probability, that is, given the ratio between inpatient admissions and DH in the original data set (equal to 1185255/187066=6.336), extracting a sample of 10\% of the total number of inpatient admissions from the subset of such admissions, and a sample of 63\% of DH admission from the respective subset.

**Results of the classification analysis**

As stated above, a classification analysis procedure was applied, using the modality of hospitalization as response variable, and all the other information in the dataset as predictive. In order to draw significant relations, we applied the CaRT algorithm with good stop criteria (maximum depth=7 levels, minimum parent-node size=60 units, minimum child-node size=30 units) and pruning option, in this way reducing the tree size as well as noise and errors\(^3\).

Following such criteria, the classification algorithm identified 79 nodes, and 40 of them can’t be further subdivided (final nodes). Unfortunately, the tree diagram (Fig. 2) is too large in order to be clearly reproduced here, and thus only a brief description of it will be exposed.

In the first level of division, the most significant explanatory variable linked to the response variable is “Main diagnosis ACC”, divided into two large subgroups, each of which includes approximately half of the sample. In particular, the classification of the 1\(^{st}\) node shows that patients with Essential hypertension, Varicose veins of the lower extremities, Other diseases of veins and lymphatics, Other circulatory diseases are admitted as inpatient hospitalization in 24\% of cases and as Day Hospital in the remaining 76\% of cases.

The 2\(^{nd}\) node, where all other cases are grouped, shows the opposite distribution: inpatient admission, 77\%, and Day Hospital, 23\%. The four segmentation levels that follow this node depend only by diagnostic motives (almost all the secondary diagnosis ACCs, but also some clarification given again by the principal diagnosis ACC), therefore appearing quite appropriate.

Instead, the classification levels that follow the 1\(^{st}\) node highlight more item that could be critical, at first the admission district (see Table 3, too): indeed, in Bari district the choice to the "day hospital" is considerably higher (85\% of cases) compared to the other Apulian districts (Brindisi, Foggia, Lecce, Taranto, Barletta-Andria-Trani), where the least costly mode of hospitalization records about 64\%. In the first district, in particular, the Day Hospital is chosen in almost 91\% of patients with a secondary diagnosis ACC as “Essential hypertension”, “Hypertension with complications and secondary hypertension”, “Heart valve disorders”, “Aortic, peripheral and visceral artery aneurysms” and “Other circulatory diseases”. Such diseases, however, are treated as DH in different ways depending on whether the hospital is private (98\% of cases), or public / ecclesiastical (83\% of cases).

\(^2\)The segmentation analysis moves from all the information available, dividing the sample, with iterative procedures based on the explanatory variables and their relationships, in aggregate (“nodes”) that are internally homogeneous against the response variable (Breiman et al., 1984). Ideally, all cases of a final node, which is a node that cannot be further divided, should have the same value as the response variable (having the highest purity), being maximally different from the cases of the other final nodes. The analysis, however, provides poor results when the distribution of the response variable is highly skewed (Fabbri, 1997).

\(^3\)The pruning method reduces the size of classification trees by removing tree sections (in a bottom-up way) which provide little power to classify instances. The second purpose of the pruning technique is to provide a better predictive accuracy by the reduction of overfitting.
The classification tree was validated using a training sample and a test sample (random sub-samples of approximately equal numbers, extracted from the sample obtained as described in section 3.2). The result of such analysis is certainly interesting, since the estimate of misclassification (ie the percentage of cases incorrectly classified) amounts to little more than 0.17, with a standard error equal to 0.001 in both training and test sample.

Indeed, the algorithm is able to achieve a percentage of correct classification equal to almost 83%, almost no differences between one mode and the other, and even between the training sample and the test (which, it should be emphasized, making the division using the rules identified by the training).

Applying the classification rules (recorded into a command script) to the whole population of hospitalizations for Cardiovascular disease, the membership of each observation to the one or the other predicted status was defined following the values of the respective independent variables. Despite the strong asymmetry of information, the ability of the method to reconstruct the choices for the entire dataset is unchanged from sample results.

**Concluding considerations**

In this paper the multivariate analysis of classification plays a role not common but very useful: the search for anomalies in the organizational appropriateness. Instead of a detailed study “case by case” of the organizational appropriateness (this task would require in-depth analysis of the SDO, excellent clinical expertise and, above all, large human resources and much time), an analysis was conducted on the results of the “average” choices, considering “appropriate” that choices made on the basis of clinical factors (including age of patients) and as “critical” those that were dependent on factors outside the clinical (eg., availability of personnel or the cost effectiveness of the hospital).

The obtained model is quite useful, as it shows a good distribution in the classification of cases “expected” according to the found rules, compared to those actually observed for both hospitalization modes, ordinary (inpatient) or daily (Day Hospital or Day Surgery).

In conclusion, the technique here proposed can be very attractive if used as a “preliminary study” to the interventions by the managers of the social and health sectors. Indeed, investigating further the health facilities whose behavior appears “quite inappropriate”, they could easily identify and solve problems of inefficiency; on the other hand, studying the characteristics of health facilities that highlight a greater appropriateness, the managers could borrow their patterns of behaviour, in order to improve also the results of the other health care facilities.

**Acknowledgements**

The Authors jointly designed and realized the work here described. However, F. D. d’Ovidio wrote the section 3, D. Leogrande the section 1 and the section 5, R. Mancarella the section 4, while D. Viola wrote the section 2.

The Authors thank the A.R.E.S.-Apulia, that furnished the data. The usual disclaimer applies.

**Bibliography**
