Ex-post Evaluation and Relative Effectiveness of Health Structures: an Overview

Giorgio Vittadini, University of Milano-Bicocca, Italy[a]

1. The Need for the Evaluation of a Changing Healthcare System

In recent years there has been increasing attention directed toward the problems inherent to Quality Control in the Healthcare Services. Several factors have brought about this increased interest in the quality of the operational aspects of the Healthcare field.

1) The citizen is more and more conscious of who, even if indirectly, "pays" in order to obtain hospital services, and therefore expects quality services; "quality" in healthcare services means the ability to satisfy specific needs, and is the result of scientific, technical and technological, organizational, procedural, and relational elements in which the human variable plays a primary role, interacting closely with the production processes.
2) The Healthcare system has been going through a transition from a system where the resources were assigned to structures with a reimbursement system that was independent of the way in which they were used, to one where such allocation is distributed according to prefixed figures for each pathology. The medium amount of resources used to cure patients having a given pathology is found by means of the Diagnostic Related Group system (DRG). Therefore public health structures depend on national and regional political decisions but become autonomous from an operational point of view and are organized using functional models in an entrepreneurial way, characterized by competition between various healthcare structures. Hospitals are encouraged to search for an organizational model that allows the reduction of costs by optimizing the use of available resources (efficient use of resources) and the increase of patient satisfaction due to optimal clinical

[a] Department of Statistics, University of Bicocca-Milan,Via Bicocca degli Arcimboldi 8, 20126 Milan, Italy,
Tel: ..++39 0264485865 Fax. ++39 02700569114
Email address: giorgio.vittadini@unimib.it.

conditions, including "supplementary" services (minimization of the gap between expectations and the perceptions - "customer satisfaction").

In this way, health structures have the responsibility of furnishing services that are valued in both quantitative and qualitative terms. Within public administration, Healthcare is currently the service with the highest investment related to quality, creating the opportunity to treat the Healthcare sector as a laboratory for quality control, in which internal activities are directed to assure that the demands of the patient are understood and satisfied.

3) The financing system of the Healthcare Sector based on reimbursement tends to sacrifice efficiency. The system, based on the DRG, has the disadvantage of inducing hospitals to place a higher priority on the economic aspects instead of the effective necessity of care (resulting in policies that are economically correct but ethically incorrect). This phenomenon leads to a selection of patients that are affected by a specific pathology associated with optimal clinical conditions: the evaluation is essential to find financing criteria complementary to DRG.
4) The actual situation is only the first step toward a more radical change in the health care system. It is moving in the direction of a welfare-mix system characterized by the joint-presence of state agents (operating with functional economic autonomy), private profit-oriented, private non-profit companies and by freedom of choice for the consumer, who is reimbursed for health services even when provided by accredited private agents (profit or non-profit (Borgonovi 2002, Moramarco 2002, Fiorentini 2002)).

   The consumers' total freedom of selection from among health structures can be intentionally restricted by rules, defined by political authorities; but also unintentionally by particular conditions such as the presence of information asymmetry. The consumer is not able to know many of the service characteristics that will be purchased, because these typical "experience goods " services do not manifest until the actual moment of supply. Supplying evaluation results to the consumer means giving information about the quality of the health structure that will be providing service.
5) In this context, the operating role of private structures in the financing of new healthcare infrastructures is increasing. The adoption of finance planning techniques, already used for the infrastructures and financing services of the public

sector, can be an effective instrument for generating financial resources and also for developing efficient risks allocation among the numerous stakeholders. The decisional models used in financial planning typically use a great number of variables, with the aim of estimating the entity and the kind of risks involved in determined initiatives, and of studying more efficient allocation, utilizing financial and contractual instruments. These models require evaluation analyses of the health system and of the hospitals in order to provide satisfactory results.

2. The Health System: ex ante and ex post Evaluation

Total Quality Management and Continuous Quality Improvement are the most widespread approaches and the most recent systems to implement and improve quality control in health structures. All these approaches, which include diverse theoretical-practical aspects, have a common basic idea: the importance of the evaluation and monitoring of results as a tool for initiating continuous improvement processes.

The most traditional methods for quality evaluation are the ex-ante methods, as accreditation of excellence methods and quality certification by the standard ISO 9000 in the fields of business and health.

The most famous example of ex ante system of accreditation of health structure is managed by the American Joint Commission on Accreditation of Healthcare Organization (JCAHO)[1]. The other ex ante evaluation method is given by the certification of health structures based on the guide lines UNI-EN- 9000-2000[2].

Accreditation and certification procedures can be seen as a benchmarking process which discovers systems, skills and technology that improve performance.

[1] The system is based on 355 Standard concerning patients and the organization and divided in some measurable elements. A team is composed by a clinician, a nurse and an administrator visiting every health structure. A score of 0.5 of 1 is assigned to each measurable element. The score of each standard is the weighted mean of the scores assigned to each measurable element. The whole score is assigned by means of an algorithm performed on the scores of the standard. The accreditation or the non-accreditation of health structure depends on the result of the survey (JCAHO 2004).

[2] The new guides lines of ISO –9000 - 2000 are divided in principles (8), basic issues (12), definition of the system of quality (87) concerning customers, personnel, continuous improvement, organizational process, output, conformity and other topics. The system of quality is formalized in a Manual of Quality which describes the procedures necessary to apply the guide lines. Some authorized certification agencies verify if the health structure applies the guidelines and can obtain the UNI-EN-ISO 9000 certification (UNI 2004).

However, in order to obtain and to evaluate the effectiveness of health structures we also need benchmarking results defined as ex post evaluation based on the quantitative measurement of inputs, outputs, outcomes and the relationships between them. Results benchmarking helps managers and clinicians to measure and to improve their performance.

Therefore in order to verify the quality of health structures, both accreditation and certification processes request an appropriate of data collection system for producing ex-post statistical data (Deming, 1986)[3]. Moreover, some benchmarking agencies measure health structures performances in different countries[4], without accreditation and certification processes. The different dimensions of efficiency, customer satisfaction and effectiveness are evaluated.

Efficiency consists of the comparison between production input and business output in terms of both quality (days, cases etc.) and monetary data by means of budget indexes or more sophisticated statistical methodologies (DEA, multilevel methods, production functions) applied to fit the health structure through combinations of optimal outputs given inputs (Gori, Grassetti, Rossi, 2001).

---

[3] The Joint Commission on Accreditation of Healthcare Organization (JCAHO) has been developing a consensus-driven common language and framework for performance measures. After having proposed the National Library of Healthcare Indicators, a sophisticated classification system and individual profiles for 225 performance measures, since 1997 it has been developing the ORYX program in order to introduce elements of ex-post evaluation in the accreditation process. Particularly it has been proposing some core measurements (22) connected to their best practices to assess their performance some non core measurements can be proposed by health structures and tested by JCAHO (JCAHO, 2004).

[4] Since 1993, Solucient's 100 Top Hospitals: Benchmarks for Success program has been publishing annual benchmarks for hospitals performance by identifying the top overall achievers in the industry. By naming the 2001 100 Top Hospitals, Solucient provides the latest snapshot of the ever-changing benchmarks set by the industry's highest performers, and once again describes what it takes to be a top performer (Solucient, 2003). CIHI (2004) is an independent, pan-Canadina, not-for- profit organization working to improve the health of Canadians and the health care system by providing quality, reliable and timely health information. Health indicators are standardized measures which compare health status and system performance and characteristics among different jurisdictions in Canada. A compilation of selected indicators measuring health status, non-medical determinant health, health system performance, and community and health system characteristics information is provided for Canada' s largest health regions, encompassing approximation of population, as well as provinces and territories (CIHI, 2004). In England, since July 2003, the Commission for Health Improvement (CHI) published the NHS star ratings, a composite performance assessment of NHS trusts in which each is awarded from zero to three stars. This document is a commentary on those star ratings, looking at the indicators used and the method of compiling the ratings, with a view to informing the future development of performance assessment in the health service. Some composite ratings of organizations give a single score for each organization, other seek to score organizations across a number of aspects of their performance (NHS 2004). AHRQ (2004) proposes some groups of core indicators for states hospitals, connected to their best and worse practices, that are: Patients Quality (related to effectiveness), Patients Safety (regarding surety conditions), Prevention Quality (avoidable admissions). They can be utilized by the hospitals for self quality improvement.

Customer satisfaction, which is a subject of the Marketing field and that is a set of models and methods used to measure consumer's satisfaction regarding aspects related to a product or a service, is now also applied in public sectors with the aim of evaluating the perceived quality from the point of view of public services users.

Effectiveness is defined as the ability to provide treatment to patients, improving health outcome and improving the ability to modify the patient's state of health (Donabedian, 1988) and provides useful information concerning the ability of medical personnel as well as managerial aspects of Health Services for example creating a system that provides assistance oriented services, the most effective allocation of resources and the evaluation of assistance costs and economic responsibility regarding hospitalization.

3. Ex-post Evaluation and Relative Effectiveness

An approach to quality problems in health structures is efficacy, defined as the ability to provide treatment to patients, improving health outcome and improving the ability to modify the patient's state of health (Donabedian 1988).

Of particular importance is the concept of relative effectiveness, i.e. the effect of hospital care on patients used to compare different healthcare institutions in terms of "healthcare outcomes".

Data collection for the "output" evaluation is easy, thanks to the discharge card; the collection of data on the "outcome" evaluation is more difficult. A "healthcare outcome" is definable as the "technical result of a diagnostic procedure or specific treatment episode" (Opit, 1993), a "result, often long term, on the state of patient well-being, generated by the delivery of a health service" (Aitkin, Longford, 1986; Goldstein, Spiegelhalter, 1996).

Healthcare outcomes are influenced by covariates concerning the "case mix" of the patients, definable as the variability of their clinical and social demographic aspects and are related to the organization, resources, facilities and other characteristics of hospitals (Zaslavsky, 2001). Therefore, to allow comparisons between healthcare institutions, relative effectiveness needs to be adjusted for patient-specific and hospital-specific variables by means of risk adjustment statistical methods. Depending on the presence of risk factors at the time of health care encounters, patients may experience different outcomes regardless of the quality of care provided by the health care organization. The

risk adjustment statistical methods identify and adjust variations in patient outcomes that stem from differences in patient characteristics (or risk factors) across health care organizations and, therefore, allow fair and accurate inter-organizational comparisons.

In the first case, utilizing an approach similar to that of Clinical Trials, we can observe the links between outcome and risk adjustment indicators for particular pathologies; in the second case, analyzed in this paper, we can construct evaluation systems of Health Structures valid at the Regional level by using the appropriate statistical and computer technology tables[5].

3.1 Risk Adjustment Methodology: Direct and Indirect Standardization

The first risk-adjustment methodology is called direct standardization (Zaslavsky 2001).

Given y_{kj} health outcome observed on the k-eth statum of the population of patients (i.e. kind of patients: surgical, medical patients) and

$$\pi_{kj} = w_{kj} / \Sigma_k w_{kj} \qquad (k=1,\ldots,q) \qquad (1)$$

proportion of the k-eth case mix characteristic in the j-eth health structure. The observed adjusted outcome is the weighted sum:

$$y_j = \Sigma_k \pi_{kj} y_{kj} \qquad (2)$$

Direct standardization has limitations. It may be impossible to calculate a standardized score for a stratum with no cases or missing cases. Furthermore, direct standardization is not adapted to adjusting simultaneously for many variables or for (continuous) variables.

An alternative approach is indirect standardization. Given the weighted sum:

$$\hat{y}_j = \Sigma_k \hat{\boldsymbol{\pi}}_{\mathbf{kj}} y_{kj} \qquad (3)$$

[5] As underlined by Goldstein and Spiegelhalter (1996): "When standard measures are available across organizations, they should be risk-adjusted to account for differences in environmental factors influencing outcomes".

where the weights $\hat{\pi}_{kj}$ are obtained from a standard theoretical population, in order to evaluate the relative effectiveness of the j-eth health structure we compare the observed adjusted outcomes (2) with the expected adjusted outcomes (3):

$$u_j^* = \hat{y}_j / y_j \quad (4)$$

Furthermore, indirect standardisation is not suitable when we have several case mix indicators or when they are not discrete.

3.2 Linear and Logistic Models

Various authors (Dubois et al., 1987) and the benchmark agencies use linear and logistic models as a method of risk-adjustment (AHRQ, 2003; CIHI, 2003; Solucient, 2004; NHS, 2004; JCAHO, 2004). With quantitative outcomes we have:

$$y_{ij} = \boldsymbol{\beta}\, \mathbf{x_{ij}} + e_{ij} \quad (5)$$

where y_{ij} is the j-eth quantitative outcome for i-eth patient, $\mathbf{x_{ij}}$ the corresponding patient characteristic(s), e_{ij} the error term, and β is a vector of coefficients. The first term of this equation captures the effects of individual characteristics **X** on outcomes, among patients of the same unit. In other cases covariance analysis is proposed:

$$y_{ij} = \boldsymbol{\beta}\, \mathbf{x_{ij}} + u_j + e_{ij} \quad (6)$$

where u_j measures the relative effectiveness of the j-eth agent. In the simplest formulation with dichotomic outcomes (such as hospital mortality risk) the logistic function models the logit of the outcome p_{ij} as a linear function of the case mix variables x_k (k=1,…,p):

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \ldots \beta_k x_{kij} = RS_{ij} \quad (7)$$

which can be exposed in terms of probability by means of the following formula:

$$p_{ij} = \frac{e^{a} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \ldots \beta_k x_{kij}}{1 + e^{a} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \ldots \beta_k x_{kij}} = \frac{e^{RS_{ij}}}{1 + e^{RS_{ij}}} \tag{8}$$

where p_{ij} is the observed probability (equal to 0 or 1) regarding the positive or negative occurrence of the dichotomic outcome.

By estimating the vector of parameter β and substituting $\hat{\beta}$ in (8) we obtain the expected probability $\hat{p}_{ij}$:

$$\hat{p}_{ij} = \frac{e^{\hat{p}_{ij}}}{1 + e^{\hat{p}_{ij}}} \tag{9}$$

Comparing $\hat{p}_{ij}$ with $\hat{p}_j$ we have an estimation of the effectiveness of the j-eth health structure for the j-eth patient. Therefore the expected value of $\hat{p}_{ij}$ is obtained as:

$$\left(\hat{p}_{ij}\right) = \sum_{i=1}^{n_j} \hat{p}_{ij} \tag{10}$$

and the correspondent value of the observed probabilities $\sum_{i=1}^{n_j} p_{ij}$ the ratio

$$\hat{u_j} = \sum_{i=1}^{nj} \hat{p}_{ij} \Big/ \Sigma\left(p_{ij}\right) \tag{11}$$

estimates of the effectiveness of the j-eth health structure.

3.3 The Multilevel Model

In many cases logistic or linear models are not adequate for calculating the relative effectiveness of health structures for several reasons:

1) A sample of agents is chosen in the agents populations.
2) The data often show highly structured hierarchies because patients or episodes of care are nested within health structures and higher-level health institutions.

3) There is no basis for the prior assumption that outcome measures can be devised along a single dimension. Measures are likely to differ for subgroups of patients within a single broad category- when the variety in numbers of secondary diagnoses and patients of different ages and sexes, and previous medical histories are all likely to influence outcomes (Normand et al., 1995).

4) The ability to risk adjust using retrospective data can be hampered by the potential endogeneity of the recorded patient-level covariates[6] and the potential for non considered covariates[7]. The characteristics of both patients and health structures need to be taken into account.

5) There are problems concerning different amounts of patient subsamples in different health structures[8].

6) The methods utilized for ranking health structures are not ideal, above all for identifying extreme hospital performance[9].

7) In effectiveness studies there are cases where the distributional hypotheses are complex: the distribution of outcomes, of random effects and of the effectiveness parameters is not normal,

8) The relationships between outcomes and case mix indicators can be nonlinear.

In these and other cases the quite extensive literature about variance and mixed effects models suggests that hierarchical models and particularly multilevel models offer solutions for studying relationships between outcomes (mortality, health, quality of life) and contextual variables in complex hierarchical structures, considering both individual and aggregate levels of analysis (Goldstein, 1995; Leyland, 1995; Goldstein, Spiegelhalter, 1996; Verbeke, Molenberghs, 2000; Carlin et al., 2001).

Particularly, in the nineties, numerous authors (Thomas et al., 1994; Normand et al., 1995; Morris, Christiansen, 1996; Goldstein, Spiegelhalter, 1996; Rice, Leyland, 1996;

[6] "It is possible that there are differential error rates associated with the ability or propensity to record information across providers "(Normand et al., 1995, p.812).

[7] "It is often possible that an important severity measure will be missing from the database, and furthermore, it is likely that the distribution of this unmeasured covariate will vary across providers. As a consequence, the magnitude of the difference between the adjusted and standardized outcomes may be exaggerated"(Normand et al., 1995, p.812).

[8] "The estimates for hospitals with relatively large numbers of patients generally will be pulled only slightly toward the group mean even if they are quite different from it. In contrast, estimates of hospitals with small numbers of patients are likely to be pulled strongly toward the group mean if they different substantially from the mean. Substantial shrinkage would be justifiable in such cases, because the raw performance estimates are bound to be imprecise"(Normand et al., 1995, p.813).

[9] Technical limitations do not permit varying hospital caseloads and, therefore, that a true one-year grafit failure rate that exceeds 25% at any hospital (Morris et al., 1996).

Leyland, Boddy, 1998; Marshall, Spiegelhalter, 2001; Dubois et al., 1987; Jencks et al., 1988; Epstein 1995; Schneider, Epstein, 1996) proposed studying "relative effectiveness" by means of the Multilevel Model (Hox, 1995)[10].

The Multilevel Model specified for the j-th outcome is:

$$y_{ij} = \boldsymbol{\beta_j}\, \mathbf{x_{ij}} + u_j + e_{ij} \qquad (12)$$

where u_j is the random coefficient interpretable as the effectiveness of hospitals with respect to outcome y_{ij} adjusted for patient characteristics made up of fixed coefficients of patient covariates.

Casual parameters are usually obtained by means of bayesian inference (Leyland, Boddy, 1998) and they can have prior distribution, above all, in the case of patients. Subsequently, particular attention must be paid to estimation parameter procedures as well as the statistical properties of the estimators with respect to single model parameters and different parameter typologies (fixed effects and variance components at different levels of hierarchical structure) as a whole.

The authors cited above have considered the Multilevel Model under the following restrictions:

a) the consideration of one outcome at a time as the response variable;
b) assumption of binomial distribution of the outcome;
c) no consideration of hospital-specific characteristics as possible covariates;
d) assumption of multinormal distribution for random disturbances and the random parameters of effectiveness (under the hypothesis that they are independent and identically distributed);
e) static models;
f) non correlation between the expected values of the patients characteristics and the effectiveness casual parameters.

These assumptions are extremely restrictive and are not sufficient for building a methodology for the evaluation of a health system. Generalizations concerning statistical

[10] "The Multilevel Model overcomes small sample problems by appropriately pooling information across institutions, introducing some bias or *shrinkage*, and providing a statistical framework that allows one to quantify and explain variability in outcomes through the investigation of institutional level covariates" (Marshall, Spiegelhalter, 2001, p.128).

theory and choice of outcomes and risk adjustment variables can be advanced (Pagano, Vittadini, 2004).

4. Methodological Generalizations

1 First of all, if we interpret outcomes as latent constructs underlying a set of observable indicators (Gertler, 1988) they can be defined in statistical terms as latent variables (Muthen, Speckart, 1985; Leyland, Goldstein, 2001).

In order to obtain solutions we avoid traditional structural equation models, which lead to indeterminacy of latent scores (Vittadini, 1989). Instead Partial least Squares and Regression Component Decomposition methods (Schöenemann, Steiger, 1976), which approximate latent variables by means of linear combinations of their indicators (Vittadini, 2001) can be utilized.

2 When the indicators are qualitative or mixed, there are various procedures of quantification in literature: in order to obtain quantified observable indicators and simultaneously their quantitative linear transformations, alternate multidimensional scaling methods, such as Alsos Princals algorithm, can be utilized (de Leeuw, van Rijckervorsel, 1980; Young, 1981).

However, this modus operandi presents some drawbacks. The methodologies of latent variables estimation should have to guarantee the availability of "objective performance measures" obtained taking into account the difficulty of a trial of tests (Vittadini, 2001; Lovaglio, 2003). The Rasch model is suitable for the estimation of latent outcomes because it allows the estimation of objective measurement of performance with the most agreeable properties (Wright, Masters, 1982).

3 Of interest is the study of the simultaneous dependency of covariates from multiple outcomes, which can also be correlated to each other (e.g. the correlation between mental and physical states of health in quality of life[11]).

[11] In this case the combination of the Multilevel Model with the Seemingly Unrelated Regression Equations has been proposed (SURE) Model (Srivastava, Giles, 1987) in a unique model, entitled the SURE Multilevel Model (Vittadini, 2001; Vittadini, Minotti, 2005). There are two reasons for this. First, the SURE Model consists of a system of simultaneous equations, equal in number to the response variables, where disturbance terms related to different individuals in the same equation are uncorrelated, but disturbance terms in separate equations are correlated. Secondly, the SURE Model allows the specification of different regressors for each outcome. For example, in the case of two different outcomes (final mental state of

4 Relative effectiveness needs to be adjusted for hospital characteristics, such as resources, organizational capacity, etc..., for comparison studies. We can introduce further equations describing hospital characteristics. That means specifying a second level equation (i.e. referring to hospitals), which expresses the random parameters u_{jv}, (j = 1, ..., p; v = 1, ..., q), as a function of hospital characteristics (Normand et al., 1997, p.812).

5 The assumption of multinormality or binomial distribution for random disturbances and random parameters of relative effectiveness is often too restrictive and above all is not respected when the indicators are qualitative or mixed or no prior information on parameter distribution is available (Longford, Lewis, 1998; Marshall, Spiegelhalter, 2001). As it is well known, in this situation the usual hypothesis of normally distributed residuals and parameters is not suitable. Therefore we can employ families of distributions other than normal[12].

6 Mixed effects models assume that random agents effects are independent and non correlated with the expected values of patients characteristics. In complex models, this assumption might not be realistic, in particular in the case of panel (longitudinal) data, where one of the levels is represented by the subject. Individual heterogeneity can be poorly modelled by the available subject-level covariates, and this fact can dramatically affect the estimation of random effects (Neuhaus, Kalbfleisch, 1998; Heagerty, Kurland, 2001). In the linear case, possible remedies are based on the use of fixed-effects modelling in place of random effects, correcting the incidental parameter problem caused by using one parameter for each subject: this is usually done by conditioning out the subject-level parameter (Verbeke, Molenberghs, 2000). In non-linear cases, fixed-effects estimation is more complicated; we cannot rely on exact conditioning to eliminate the subject-specific parameter, as this is possible only in a generalized linear model with canonical link functions[13].

7 Casual effects models are proposed when the patients variability "within" is hypothesized stochastic and therefore the patients parameters are random parameters.

---

health and final physical state of health), we can indicate the initial state of health (which is different for each outcome), from among the various explicative variables.

[12] As the "exponential power distribution" (Subbotin, 1923; Box, Tiao, 1992, Vittadini, Minotti, 2005), the exponential distribution (Yang, 2001), the politomic hierarchical distribution (Daniels, Gatsonis, 1997), prior non informative distributions for the effectiveness stochastic parameters (Vittadini, Sanarico,).

[13] Instead, we have to resort to approximate conditioning, as shown in Bellio and Sartori (2003) for the case of binary data.

8 Interactions between effectiveness parameters and patient characteristics can also be proposed when effectiveness changes for different patient groups inside hospitals (Leyland, Boddy 1998).

9 To evaluate the effectiveness of territorial health care, relative to specific pathologies, it could be interesting to take in consideration the time trends of the hospital services demanded, as they can be viewed as substitutive services. Conditionally on the administrative data availability, dynamic panel data models and longitudinal multilevel models can be applied in this context (Verbeek, Molenberghs, 2000), quite often non linear models (Ann Gilligan et al., 2002; Yang, 2001).

When the data typically available in health analysis has an observational nature (as in the case of administrative data) most of the methods developed for analysis within experimental settings may give very misleading results when applied to observational settings, and random effects models are no exception.

We can compare studies conducted with the multilevel model based on observational and epidemiological data with parametric data (Longford, Lewis, 1988), non parametric with mixed and individual data (Goldstein 1995, Goldstein, Spiegelhalter 1996, Leyland 1995, Carlin ed al. 1995).

10 Generally the linear assumption of a model is reasonable in the first request, but the utilization of hierarchical non linear models (Burns et al.1997; Yang 2001; Ann Gilinger et al., 2001) results as more opportune and appropriate, especially when dealing with units coming from a hierarchical sampling, with variables of different nature (qualitative and quantitative), often highly correlated, in the presence of missing values (Rubin, 1984) or outliers of non-linear multivariate regression (Friedman, 1991).

11 Which problems do the linear models and, in particular the multilevel models involve, when they are used with a large base data as occurs with administrative data? Besides problems connected with the accuracy of data (i.e. coding accuracy, timing of diagnoses uncertainty etc. (Damberg et al., 1998)), there is a relevant methodological problem. It is known that when there are large data sets the significance tests associated with linear models refuse the null hypothesis in all cases. In fact, the sample size influences the results, and beyond a certain threshold is the only determining factor of the test. Every explicative variable seems to be significant for explaining the outcomes and this result is particularly misleading for the topics mentioned above (Vroman Battle, Rakow, 1993). Therefore we need appropriate testing procedures able to verify

hypotheses regarding the significance of explicative variables in samples drawn from the population associated with administrative data. In general terms, we must devise inference methods in heterogeneous samples collected from large data sets (Duncan et al., 1998).

12 The estimation of complex random effects models needs suitable computational methods. Among the recent developments is the possibility to resort to simulation methods (McCulloch, Searle, 2000; Contoyannis et al., 2001) or to modern numerical techniques (Skaug, 2002).

13 Recently, this use of risk adjusted comparisons for benchmarking health structures has been strongly criticized (Lilford et al., 2004). In particular, it has been stated that: "The sensitivity of an institution's position in league tables to the method of risk adjustment used suggests that comparisons of outcomes are unlikely to tell us about the quality of care" and "Outcome is neither a sensitive nor a specific marker for quality of care" (Lilford et al., 2004). Therefore it has been suggested that: "The agencies should facilitate the development and dissemination of a database for best practice and improvement based on the results for primary and secondary research" (Lilford et al., 2004). In this direction the use of administrative data, used by payers to pay bills and manage operations, can be very useful[14].

5. Outcome and Risk-Adjustment Generalizations

In this context, the debate regarding types of outcomes (clinical, quality of life, context) and risk adjustment variables is of particular importance.

Different types of generalizations can be made regarding outcome, explicative variables and data sets.

Death rate is the most utilized health outcome in effectiveness studies. However there are many doubts regarding its use[15]. Moreover the list of variables involved in case

[14] In fact: "These data are typically computerized, making it easy to collect and use large quantities of information… Administrative data have been used to examine geographic variation in utilization of surgical and medical procedures, monitor the use of health services, assess the effects of a policy change on health expenditures, evaluate the relationships between hospitality death rates and hospital characteristics" (Damberg et al., 1998).

[15] "The relationship of death rates to quality of care remains controversial and unproven. Unless finding are adjusted for patient characteristics, conclusions about quality based on an evaluation of [patient] outcomes

mix studies cannot be complete and this can generate a bias in outcomes evaluation, because small differences in case mix can have significant effects on outcome measurements[16]. Therefore:

a) In many epidemiological studies we utilize clinical outcomes and risk adjustment differently from traditional mortality rates. In fact, such clinical "outcomes" describe the quality of the treatment used in the various pathological conditions. Moreover numerous case mix indicators have been developed specifically for comparing hospitals or large patient groups[17]. They are obtained from clinical data which represent the most important source of information about the details of diagnosis and treatment, patient risk factors, and the clinical outcomes of care. But there are many problems regarding their use. Sometimes medical records do not contain a unique individual identifier for patients and each patient can have multiple medical records that are the result of care provided by different doctors and hospitals. Moreover they do not have standard format or procedures for recording patient information and are affected by random fluctuation of outcomes from year to year. They are contained in medical records, typically handwritten, and therefore the abstraction process is very costly and time-consuming (Damberg et al., 1998). Often case mix indicators do not take into account patients' and health structures' characteristics[18].
b) The international agencies, which benchmark hospitals, propose proxy of clinical outcomes and risk adjustment indicators (CIHI, 2004, CHI, 2004, JCAHO, 2004, NHS, 2004). They are Process measures, Outcomes measures, and Sentinel events. Process measures, expressed in terms of a percentage, or rate, describe how often a series of activities, actions, or steps are carried out (for example, a treatment

---

may be erroneous." (Iezzoni et al., 1996, p.1379). "Mortality may depend more on the mix of patients reaching hospital in the first place, rather than the quality of care given once admitted. In hospital mortality in particular may be prone to bias and manipulation" (Goldstein, Spiegelhalter, 1996, p.399).

[16] Measures of severity of illness at admission to hospital have been criticized for not fully taking into account known discrepancies in outcomes associated with social background and other factors (Goldstein, Spiegelhalter, 1996, p.399). "The majority of variation in annual hospital death rates for the four conditions studied (stroke, pneumonia, myocardial infarction, and congestive heart failure) is chance variability that results from the relatively small numbers of patients treated in most hospitals in a year… Risk adjustment methods do not show whether the unexplained difference in mortality rates results from differences in effectiveness of care or unmeasured differences in patient risk at the time of admission" (Jencks et al, 1994, p.3611).

[17] Many are commercial, proprietary products, marketed to hospitals, government officials, legislators, and business leaders (Iezzoni et al., 1996, 1379).

[18] Leyland, Boddy, pp. 537, 555.

such as aspirin at arrival) in a patient population over a set time period. The numerator is a count of the number of patients in the measure population who had the treatment or experienced the event. The denominator is the total number of patients for whom the treatment or event was appropriate. Outcomes measures expressed in terms of a percentage or rate, describe the results of the performance of a function or process (for example, vaginal tears during delivery) in a patient population over a set period of time. The numerator is a count of the number of patients in the measure population who had the treatment or experience the event. The denominator is the total number of patients for whom the treatment or event was appropriate.

Sentinel event is an unexpected occurrence, involving death or serious physical or psychological injury, or the risk thereof (JCAHO, 2004) [19]. The types of measures are very useful in order to improve effectiveness, because they are linked to best practices in health care. However it is very expensive and very difficult to collect these effectiveness indicators: the same difficulties described for clinical indicators are valid for proxies of clinical indicators. Moreover a general quality system must be built in order to collect them.

c) Quality of life indicators refer to the general condition of health of the patent or of context, and describe the conditions by means it has been distributed the service. We can utilize quality of life outcomes: in particular SF-36 and SF-12 scales (Ware et al., 1986; Ware et al. 1992a; Ware et al., 1992b) are measure instruments based on patients questionnaires in order to obtain information about physical and mental health. The functional state can be measured, as an example, through technical scale FIM, than in rehabilitation measure the individual functional independence degree.

However this quality data is collected by means of Survey Data. Surveys can be costly to perform and should be performed only when an important need arises. Often surveys are conducted to collect data that are not available through existing data sources. The majority of surveys are designed as cross-sectional studies. As a result, they allow you to evaluate only what is occurring at a given point in time as

[19] Serious injury specifically includes loss of limb of function. The event is called "sentinel" because it should send a signal or sound a warning that requires immediate attention.

opposed to being able to examine the behavior or event over time. The results derived from a survey may be biased for various reasons (Damberg et al., 1998) [20].

d) Instead of clinical outcomes or quality of life outcomes (Tesio, 2003; Lovaglio, 2003; Pagano, Vittadini, 2004), we can use context outcomes which are less precise but more easily obtained from administrative data. Some methods such as disease staging or APR DRG transform the DRG administrative data using indicators as urgency of hospitalization or comorbities to obtain ordinal variables where the order of modalities is given by seriousness of diseases of patients in charge (Iezzoni et al., 1996). Administrative data have the advantage of capturing information on a large number of individuals, systemwide; however, these data currently have certain limitations that must be addressed to develop a high-quality, integrated data system. For many procedures and diagnoses, different diagnosis and procedure codes can be used to describe the same event. The accuracy of diagnostic coding can vary substantially across hospitals and physicians[21].

[20] "For example, the sample drawn from the population may not be representative or the survey may suffer from low response rates or poor recall on the part of respondents...Surveys do not contain linking variables as a general rule. It may not be possible to combine survey information with other information on the respondent… There can be significant delays in the release of the data after the survey is done… Such delays affect the timeliness of the information" (Damberg et al., 1998, pp.68-69).

[21] "Diagnosis codes tend to encompass broad ranges of disease severity and therefore may mask important clinical subgroups that differ in their expected response to treatment and do not allow direct determination of the patient's severity of illness. To assess severity of illness often requires obtaining data from the patient's medical record. This is difficult to do, since it involves manually abstracting information from the medical record... Since diagnostic information appears on administrative forms, it is impossible to distinguish complications that arise during the course of treatment from comorbid conditions that are present at the time the patient is admitted to the hospital" (Damberg et al., 1998, pp.52-53).

References

Aitkin, M, Longford, N. (1986), "Statistical Modelling Issues in School Effectiveness Studies", *J. Roy. Statistical Society*, **149**(1), 1-43.

AHRQ (2003), Agency for Healthcare Research and Quality, Guide to Inpatient Quality Indicators, US Department of Health and Human Services, Rockville, www.ahrq.gov/dat/hcup.

Ann Gilligan, M., Kneusel, R.T., Hoffmann, R.G., Greer, A.L., Nattinger, A.B. *(*2002), "Persistent Differences in Sociodemographic Determinants of Breast Conserving Treatment Despite Overall Increased Adoption", *Medical Care*, **40**(3), 181-189.

Borgonovi, E.(2002), Che cos'è il Welfare Mix?, in: G.Cittadini (eds.), *Liberi di scegliere*, Etas, Milano.

Carlin, J.B., Wolfe, R., Brown, C.H., Gelman, A. (2001), "A Case Study on the Choice, Interpretation and Checking of Multilevel Models for Longitudinal Binary Outcomes", *Biostatistics*, **2**, 397-416.

CIHI, Canadian Institute for Health Information, http://secure.cihi.ca/cihiweb/splash.html.

Contoyannis, P, Jones A.M., Leon-Gonzalez R. (2001), "Using Simulation-based Inference with Panel Data in Health Economics", Working paper, University of York, Department of Economics and Related Studies, available at http://www.york.ac.uk/res/herc/yshe.

Damberg C., Kerr E.A., Mc Glynn E.A. (1998), Description of Data Sources and Related Issues, in Mc Glynn E.A., Damberg C., Kerr E.A., Brook, R.A. (eds.), *Health Information Systems. Design Issues and Analytical Application,* RAND Health Corporation, **5**, 43-76.

de Leeuw, J., van Rijckervorsel, J. (1980), "Homals and Princals, Some Generalizations of Components Analysis", *Data Analysis and Informatics*, **2**, 231-241.

Deming, W.E. (1986), "Out of the Crisis", MIT/CAES.

Donabedian, A. (1988), "The Quality of Care. How can it be assessed?", *JAMA*, **260**(12), 1743-1748.

Dubois, R.W., Brook, R.H., Rogers, W.H. (1987), "Adjusted Hospital Death Rates: Potential Screen for Quality of Medical Care", *Am J Public Health,* **77**, 1162-1167.

Duncan, C., Jones, K., Moon G. (1998), "Context, Composition and Heterogeneity: Using Multilevel Models in Health Research", *Soc Sci Med*, **46**, 97-117.

Epstein, A. (1995), "Performance Reports on Quality-Prototypes, Problems and Prospects", *N Engl J Med,* **333**, 57-61.

Friedman, J.H. (1991), "Multivariate Adaptive Regression Splines", *Ann. Stat.,* **19**(1), 1-141.

Gertler, P.J. (1988), "A Latent-Variable Model of Quality Determination", *J Bus Econ Statist*, **9**(3), 241-252.

Goldstein, H. (1995), *Multilevel statistical models*, (2nd Ed.), Edward Arnold, Wiley, New York.

Goldstein, H., Spiegelhalter, D.J. (1996), "League Table and their Limitations: Statistical Issues in comparisons of Institutional Performances (with discussion)", *J. Roy. Statistical Society,* **159**(5), 385-443.

Gori, E., Grassetti, L., Rossi, C. (2001), "La valutazione dell'efficienza delle strutture ospedaliere: il caso della regione Lombardia", *Atti del Convegno intermedio della Sis.*

Heagerty, P.J., Kurland, B.F. (2001), "Misspecified Maximum Likelihood Estimates and Generalised Linear Mixed Models", *Biometrika*, **88**, 973-985.

Hox, J.J. (1995), *Applied Multilevel Analysis,* TT-Publikaties, Amsterdam.

Iezzoni, L.I., Ash, A.S., Shwarts, M., Daley, J., Hughes, J.S., Mackiernan Y.D. (1996), "Judging Hospitals by Severity-Adjusted Mortality Rates: the Influence of the Severity-Adjustement Method", *Am J Public Health,* **86**(10), 1379-1387.

JCAHO (2004), Joint Commission on Accreditation of Healthcare Organizations, www.jcaho.com/accreditated+organizations/behavioural+health+care/oryx/index.htm.
Jencks SF, Daley J, Draper D, Thomas N, Lenhart G, Walker J. (1988), "Interpreting Hospital Mortality Data", *JAMA*, **260**(24), 3611-3616.
Leyland, A.H. (1995), "Examining the relationship between length of stay and readmission rates for selected diagnoses in Scottish hospitals", *IMA J Math Appl Med Biol,* **12**, 175-184.
Leyland, A.H., Boddy, F.A. (1998), "League tables and acute myocardial infarction", *Lancet*, **351**, 555-558.
Leyland, A. H., Goldstein, H. (eds.) (2001), *Multilevel Models of Health Statistics,* J. Wiley, London.
Lilford, R., Mohammed, M.A., Spiegelhalter, D.J., Thomson, R. (2004), "Use and Misuse of Process and Outcome Data in Managing Performance of Acute Medical Care: Avoiding Institutional Stigma", *The Lancet*, **364**, 1147-1154.
Lombardo, R. (2003), "Singly and Doubled Ordered Non Symmetric Correspondence Analysis", Submitted.
Longford, I., Lewis, (1998), "Outliers in Multivariate Data", *J. Roy. Statistical Society – Series A*, **161**, 121-160.
Lovaglio, P.G. (2003), "The Estimate of Customer Satisfaction in a Reduced Rank Regression Framework", *Total Quality Management*, **16**, 33-44.
Marshall, E.C., Spiegelhalter, D.J. (2001), "Institutional Performance", Goldstein H., Leyland A.H. (eds.), *Multilevel Modelling of Health Statistics,* Wiley, Chichester, 127-142.
McCulloch, C. E., Searle, R.E. (2000), *Generalized, Linea, and Mixed Models*, J. Wiley, New York.
Morris, C.N., Cristiansen, C.L. (1996), "Hierarchical Models for Ranking and for Identifying Extremes", with application, in: Bernardo, J.O., Berger Dawid, A.P. (eds.), *AFM In Bayesian Statistic 5,* Oxford University press, 277-297.
Neuhaus, J.M., and Kalbfleisch, J.D. (1998), "Between- and Within-Cluster Covariate Effects in the Analysis of Clustered Data", *Biometrics*, **54**, 638-645.
NHS (2004), National Health Service, CHI, Commission for Health Improvement, A Commentary on Star Ratings 2002-2003, www.chi.nhs.uk./ratings.
Normand, S.L.T., Glickman, M.E., Ryan, T.J. (1995), "Modelling Mortality Rates for Elderly Heart Attack Patients: Profiling Hospitals", in: Gatsonis, C., et al. (eds.), *The Co-Operative Cardiovascular Project. Case Studies in Bayesian Statistics*, Springer Verlag, New York, 435-456.
Opit, L.J. (1993), "The Measurement of Health Service Outcomes", *Oxford Textbook of Health Care*, **10**, OLJ, London.
Pagano, A., Vittadini, G. (eds.) (2004), *Qualità e valutazione delle strutture sanitarie*, Etas, Milano.
Rice, N, Leyland, A (1996), "Multilevel Models: Applications to Health Data", *Health Serv Res Policy*, **3,** 154-164.
Rubin, D.B. (1984), "Bayesianly Jusfigiable and Relevant Frequency Calculations for the Applied Statistician", *The Annals of Statistics*, **12**(4), 1151-1172.
Schneider, E.C., Epstein, A.M. (1996), "Influence of Cardiac-Surgery Performance Reports on Referral Practices and Access to Care", *New Eng J Med*, **335**(4), 251-256.
Schönemann, P., Steiger, J. (1976), "Regression Component Analysis", *Br J Math Stat Psych*, **29**, 175-189.

Skaug, H.J. (2002), "Automatic differentiation to facilitate maximum likelihood estimation in nonlinear random effects models", *Journal of Computational and Graphical Statistics,* **11**, 458-470.

SOLUCIENT LLC, **(**2003), *Solucient's 100 Top Benchmark Hospitals*, www.solucient.com.

Srivastava, V.K., Giles, D.E.A. (1987), *Seemigly Unrelated Regression Equations Models*, Mercel Dekker, New York.

Tesio, L., (2003), "Measuring Behaviours and Perceptions: Rasch Analysis as a Tool for Rehabilitation Research", *J Rehabil Med.*, **35**, 105-15.

Thomas, N., Longford, N.T., Rolph J.E. (1994), "Empirical Bayes Methods for Estimating Hospital Specific Mortality Rates" *Stat Med,* **13**, 889-903.

Verbeke, G., Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, Sprinter Verlag, New York.

Vittadini, G. (1989), "Indeterminacy Problems in the Lisrel Model", *Multivariate Behavioural Research*, **24**, 397-414.

Vittadini, G. (2001), "On the use of multivariate regression models in the context of multilevel analysis", in: Borra, S., et al. (eds.), *Advances in Classification and Data Analysis*, Springer, 225-232.

Vittadini, G., Minotti, S. (2005), "A methodology for measuring the relative effectiveness of healthcare services", *IMA*, http://imaman.oxfordjournals.org/papbyrecent.dtl.

Vroman Battle, M., Rakow, E.A. (1993), "Zen and the Art of Reporting Differences in Data that are not Statistical Significant", *IEEE Transactions on Professional Communication*, **36**(2), 75-80.

Ware, J.E., Kosinski, M., Bayliss, M.S., Mc Horney, C.A., Rogers, W.H., Raczek, A. (1986), "Comparison of Methods for Scoring and Statistical Analysis of SF-36 Health Profiles and Summary Measures: Summary of Results from the Medical Outcomes Study", *Medical Care*, **33**, AS264-279.

Ware, J.E., Sherbourne, C.D. (1992a), "The Mos 36-Item Short-Form Health Survey (SF36): Conceptual Framework and item selection", *Medical Care*, **30**, 473-483.

Ware, J.E., Sherbourne, C.D., Davies, A.R. (1992b), "Developing and Testing the Mos 20-Item Short-Form Health Survey: a General Population Application", Stewart A.L, Ware JE (eds.).

Wright, B.D., Masters, G.N. (1982), *Rating Scale Analysis, Rasch Measurement*, MESA, Chicago.

Yang, M. (2001), "Multinomial Regression", in: Leyland A. and Goldstein H., *Multilevel Models of Health Statistics,* J. Wiley, London.

Young, F. (1981), "Quantitative Analysis of Qualitative Data", *Psychometrika*, **46**, 357-388.

Zaslavsky, A. (2001), "Statistical Issues in reporting Quality Data: Small Samples and Casemix Variation", *Int J Qual Health Care*, **13**(6), 481-488.