Methods for evaluating activities within the Faculty of Medicine and Surgery: case study of the "Polo San Luigi Gonzaga" in Orbassano

Giuseppe Costa[1], Dario Gregori[1], Paola Berchialla[1], Pasquale Pagliaro[2] and the Working Group for the Development of the "Polo San Luigi Gonzaga"[3]*

[1] Department of Public Health and Microbiology, University of Turin, Italy
[2] Department of Clinical and Biological Sciences, University of Turin, Italy
[3] Faculty of Medicine and Surgery, University of Turin, Italy

Corresponding Author:
Prof. Dario Gregori
Dipartimento di Sanità Pubblica e Microbiologia
Via Santena 5bis, 10126 Torino, Italy
E-mail: dario.gregori@unito.it

Summary
An evaluation system in the setting of a medical faculty focuses on medical care, research and teaching activities, and is intended to be used as a decision-making tool for assigning resources among groups and for fostering the development of the faculty. In short, it serves as a tool for evaluating and stimulating the activities of groups within the medical faculty and the university.
This article begins by presenting a set of evaluation criteria, their conceptual justification and the operational indicators used to monitor them. This is followed by a description of the calculation method (based on the TOPSIS multicriteria decision-making model) used for ranking workgroups' activities. Boxes, illustrating the application of the evaluation system at the "Polo San Luigi Gonzaga" in Orbassano (Faculty of Medicine and Surgery, University of Turin) accompanies each section of the description.

KEY WORDS: *multicriteria decision making, medical faculty evaluation, resource allocation, capacity development*.

Introduction

The complexity of medical faculties, which have a triple vocation (medical care, teaching and research) and in which there co-exist workgroups with heterogeneous interests, ranging from biological to chemical, medical and surgical, renders efforts to evaluate the activity of their members both extremely important and also particularly difficult.
Much work has recently been done with regard to the evaluation of some specific aspects, primarily the quality of teaching (1-7); quality of care has also been the focus of considerable attention, albeit often limited to specific medical care sectors (8-11). General evaluation criteria have been developed (12-22), but also harshly criticised (23).
The aim of this work is to present and discuss critically the solution adopted by the teaching hospital facility (known as the "Polo San Luigi Gonzaga" of Orbassano, hereinafter "Polo") of the Faculty of Medicine and Surgery of the University of Turin, in order to develop a system for evaluating medical

* The authors are grateful to the editor-in-chief for his helpful comments and suggestions. The authors retain full responsibility for the final version.

care, research and teaching activities carried out by the groups belonging to it. This evaluation system is intended to be used primarily as a decision-making tool for assigning resources among groups and for fostering the development of the "Polo": a tool for evaluating its groups' activities.
The article begins by presenting the evaluation criteria used, their conceptual justification and the operational indicators used to monitor them. Then, a conceptual framework of multicriteria decision making is presented, and the calculation method used for ranking university groups is described. Each part of the description is accompanied by a box, which illustrates the application of the evaluation process to the "Polo". In the final section, the article discusses the validity of the evaluation system, also considering data relating to the "Polo". A worksheet has been made available on this journal's website, providing practical illustrations of the computations and showing how data should be organised before proceeding with the evaluation and the production of the final rankings.

Subjects to which the evaluation is applied

The evaluation process is not applied to single university professors, or to specific scientific disciplines, but to the operational units, or groups, that carry out the faculty's teaching, research and medical care activities. These operational units generally correspond directly (*i*) to specific sectors of the basic sciences, hereinafter scientific-disciplinary (SD) sectors, and (*ii*), with regard to medical care disciplines, to the medical care structures defined in the hospital deed (as indicated in the classification used in hospital and regional information systems).
The first distinction between basic science and medical care disciplines is based on the activities characterising the two areas, which constitute the basis of the evaluation process. Since the activities carried out in the basic science and medical areas do not coincide completely (only research and teaching in the former, also medical care in the latter), it is difficult to compare them. It is therefore reasonable to divide the evaluation process into two parts, splitting the resources to be allocated (typically, the staffing budget) between these two areas at the start of it. To do this, criteria are needed for this initial allocation of resources, after which the evaluation system must provide a ranking of workgroups for the further allocation of resources within each area.
To decide how to allocate the resources between the basic science and medical care areas, the most feasible approach is perhaps to evaluate the quantity of professional resources absorbed by the functions that the two areas have in common, i.e. teaching and research. For the former, this quantity can be estimated on the basis of the number of university formative credits (CFUs, including credits for training), for the degree course in medicine and surgery, that are assigned to each of the two areas. For research, on the other hand, there is no simple, prescriptive criterion that can be applied to establish the quantity of resources likely to be absorbed in the two areas; the quickest way, therefore, is perhaps to take the current split of staff between the two areas (often based on the division of CFUs) as a good indicator of the research potential of each of them.
In the medical care area, university professors working in the hospital that hosts the Faculty of Medicine and Surgery are required to devote half of their working time to hospital medical care duties. In the light of this, the contribution, in terms of person-hours, that each staff member is able to make to teaching and research activities corresponds to a maximum of 50% of their working time, depending on the agreement reached with the hospital (i.e., the amount of reimbursement and support for teaching and research the operational group is allowed).
In addition, remaining within the medical care area, the weight of the care function differs in the non-surgical and surgical sub-areas; in the latter, for example, parameters for evaluating technical ability are pre-eminent and may prevail over scientific ones. In this case, too, the issue can be resolved by applying beforehand two different evaluation systems to the two sub-areas, splitting the budgetary resources at the beginning of the evaluation process, applying the same criteria mentioned above, i.e., considering the CFUs assigned to the relative SD sector. Nonetheless, this further splitting is barely feasible even in medium-small sized faculties, due to the typically small volume of the available budget, which cannot undergo two subsequent divisions, the first between the basic science and medical care areas and the second between surgical and non-

surgical sub-areas. In this case, if the decision is taken to maintain the division at the top level, between the basic science and medical care areas, it is necessary to work on an evaluation algorithm to attribute different weightings of the three functions (research, teaching and care delivery) for the two sub-areas.

Box 1
Definition of basic science and medical care in the case of the "Polo San Luigi"

With regard to teaching duties, faculty rules stipulate that 40% (126 CFUs) of the total amount of CFUs must be covered by the basic science area and the remaining 60% (189 CFUs) by the medical care area. These two areas fulfil these teaching duties with, respectively, 39% and 61% of the professors belonging to the "Polo".
In the light of this, it is also be possible (if we assume that teaching and research are the only two functions to be taken into consideration) to arrive at a reliable estimate of the level of commitment to research: i.e. at an a priori subdivision of the budgetary resources, allocating about 40% and 60% respectively.
However, the agreement reached between the university and the hospital where the "Polo" is based obliges university staff specialising in the medical care area to devote 50% of their time to medical care activities. This means that whereas the 126 CFUs that must be covered by the basic science area are guaranteed, as mentioned, by 39% of the "Polo" professors belonging to that area, each of whom contributes a professor-year for these purely university functions, the 189 CFUs for the clinical area are the responsibility of 60% of the "Polo" professors belonging to this area, who instead are able to fulfil this university function in a measure of just 0.5 professor-years each.
Therefore, since the professors from the medical care area can devote only 50% of their time to teaching and research activities, their CFUs should be worth double when the budget is allocated, thus taking the total to a nominal level of 378 credits. At this point, this amount would be equal to 75% of the total CFUs.
The division of the budget between the two areas should thus correspond to 25% for the basic science area and 75% for the care area. However, since, according to an agreement between the hospital and the university, the professors carrying out care activities receive an integrative payment, and since the agreement also allows hospital staff to be used for teaching purposes (mainly for clinical tutoring), it is agreed that a fair division of the budget would be around 33% for the basic sciences area and 67% for the medical care area.
With regard to the distinctions between sub-areas within the care area, the surgical sub-area is evaluated on a scale of priorities that assigns 40% priority to care, 35% to teaching, and 25% to research activities, while the non-surgical sub-area is evaluated on a scale that assigns 30% priority to care, 35% to teaching, and 35% to research activities.

Evaluation of the teaching function

The teaching aspect of the academic role can be evaluated through the application of various criteria, each measurable using specific indicators:

1. *saturation of compulsory teaching hours*, derived from the comparison between the designated hours and those actually spent on primary teaching tasks by each group member;
2. *other teaching-related activities* (relating to specialisations, doctorates, optional teaching activities, apprenticeships, institutional activities), for which the workload is estimated with various levels of precision;
3. *student satisfaction with teaching*, derived from the conduction of surveys among students. The indicators used are based on the scores obtained by the group's members. These scores, which correspond to students' evaluations of specific aspects of teaching, can, for a given course, be averaged across the members of each group, and standardised to the average of evaluations expressed for other courses in the same year (providing the necessary variables are available for all courses attended and that they can be drawn from survey systems of guaranteed quality and completeness).

Box 2
Evaluation of teaching at the "Polo San Luigi"

The first criterion, saturation of compulsory teaching hours, is measured as the ratio between the hours assigned to primary teaching tasks (120 hours for *ordinary and associate professors, and 90 hours for researchers) and the total number of hours spent on such activities. The number of hours is self-certified and the total cannot exceed by more than 10% the total number of hours corresponding to the number of assigned CFUs. All teaching hours (100%) devoted to the specialty degree course in medicine and surgery count as primary teaching tasks, as opposed to 50% of those spent teaching on degree courses relative to health professions pertaining to the same Faculty of Medicine and Surgery. The assignment of lower priority, in this second case, is due to the availability of integrative teaching provided by regional health service professionals, which is funded by the Piedmont Region. Hours spent on teaching activities outside the Faculty of Medicine and Surgery are not considered in the estimate, as they do not concern the budgetary resources that the evaluation system intends to allocate.*

The second criterion, which evaluates other teaching-related work, is measured through other indicators:

- *hours spent in schools of specialisation, on the coordination of doctorate work, and on teaching on elective courses (ECT) are recognised, for the purposes of the evaluation, up to a maximum of 50 hours per activity per person for each item; this is because the volume of such teaching activities is not deemed to be self-certifiable in a reproducible manner;*
- *apprenticeship hours are not counted since a reproducible estimate is not available;*
- *hours spent on institutional activities (chairmanship of faculty and degree course councils, administration of schools of specialisation, doctorate coordination) are estimated at around 20 for each activity, up to a maximum total of 60 for all the items.*

The third criterion, the students' satisfaction with the teaching received, is measured using an indicator summarising the degree of satisfaction with various specific aspects of each course, standardised according to the average of evaluations expressed for courses of the same year, providing the necessary variables are available for all courses attended and can be obtained from identification systems which are monitored for quality and completeness. Currently, the systems used for surveying students' satisfaction do not meet these requirements.

Evaluation of the research function

This aspect of the academic role may be evaluated on the basis of two different elements, both measurable using simple indicators:

1. *the quality of scientific production*, usually measured by impact factor;
2. *the ability to obtain funds for the department and the university*, measured on the basis of indicators of the ability to obtain university (COFIN – Co-Financing Scheme of the Italian Ministry of University and Research) and European funding, and indicators of the contribution of each SD sector to the university's overall ranking at national level (CIVR – Italian University System Evaluation Committee).

Box 3
Evaluation of research at the "Polo San Luigi"

Scientific productivity is measured through the average impact factor (IF) of each group, based on the IF derived from 10 papers presented by each professor or researcher in the previous five years. The average corresponds to the sum of the IFs of the group's participants divided by the number of participants. In particular, the formula adopted for calculating the IFs is adjusted by ISI class, where ISIIFScore=8(A1-B1)/(A1-1)+2, in which A1 is the number of journals in the given ISI class and B1 is the rank occupied by the journal in the same ISI class.

To illustrate the computation, let us consider how we would compute the ISIIFScore for the journals "Circulation Research" and "Basic Research in Cardiology". Circulation Research is ranked no. 2 (=B1) in IF among 71 (=A1) journals of the CARDIAC & CARDIOVASCULAR SYSTEMS ISI class, as provided by the Journal Citation Report 2004. Thus, according to the ISIIFScore formula, its adjusted IF is 9.886. On the other hand, Basic Research in Cardiology, which is ranked no. 14 (=B1) in IF in the same category (A1=71), has an adjusted IF of 8.514.

The ability to obtain investments for research is measured by two additional indicators: the COFIN factor (which measures the number of COFIN and European grants obtained in the past five years) and the CIVR factor (which measures the number of papers included in the CIVR's A and B lists). The vari-

ables included in these indicators do not present particular difficulties as regards self-certification, completeness or data quality.

Evaluation of the medical care function

The evaluation process takes the medical care function into consideration both because medical care activities can be an institutional task of academic staff (both a direct task, relative to the care of patients, and an important teaching instrument), and also because they reduce the time available for research work (in this regard they may place the academic staff involved at a disadvantage compared to researchers who do not also have medical care duties to fulfil).
Medical care activity may be examined using various methods, whose nature and importance in the evaluation are the subject of some debate; Box 4 summarises this debate in relation to the academic staff from the medical care area at the "Polo San Luigi".

Box 4
Debate over the evaluation of medical care: the corporate (hospital trust) and academic (university) points of view

Two different aspects of care activity have been considered: the administrative side and the actual delivery of care services. Evaluation of these aspects may differ according to whether it is the academic or the corporate interest that is taken in consideration.
Evaluation of the administrative activity of the health unit is relevant both to the university and to the hospital. However, whereas the hospital may be interested in the way the administrative side is managed (for example, in the management and development of the assigned human resources), the university is more interested, on the one hand, in the potential and opportunities generated by the administration of the unit as a pre-condition for steering the latter's activity towards the other university missions of training and research, and, on the other, in ensuring work loads compatible with research needs (work loads may be considered proportional to the unit's complexity).
Evaluation of administrative activity takes into account both the nature (i.e. the level of responsibility and decision-making autonomy in programming and management) and the complexity (i.e. of the resources to be managed) of the unit to be administered. With regard to this latter aspect, it is agreed that human resources are the key factor to consider, and also the easiest to measure.
Other possibly pertinent items, such as the level of investments, particularly in advanced technologies, seem to be influenced more by corporate than by university choices. Similarly, possible corporate judgements on the performance of the university unit, mainly with regard to its capacity to introduce organisational innovations, may be derived from findings of the University Evaluation Committee's appraisals of innovation stimulation projects. However, since the actual procedure of defining projects and the relative objectives, indicators and standards is not yet regulated by the university and hospital acting in concert, effective quantitative measurement of this criterion cannot be guaranteed.
Evaluation of the delivery of medical care services, which are determined by the nature of the unit, is also of interest both to the hospital and to the university, although the two have different perspectives.
The hospital focuses mainly on the efficiency criterion, i.e. on the relationship between the amount of resources spent on care activity and the level of care provided, adjusted (of course) for the amount of service requests, the level of service that the unit is asked to provide (in accordance with regional and corporate programmes), and finally, although not always, the level of satisfaction expressed by the patients.
On the contrary, the university has no specific interest in efficiency, management efficacy, or patient satisfaction (other than complying with the indications given by the hospital with which the university has a formal agreement, and providing such indications are not in conflict with its primary targets, namely its teaching and research activities).
Again, it has been emphasised that the hospital's rating of the performance of the unit reflects the hospital's own contribution to it, and in this sense it should be included in the evaluation process. Nevertheless, there is still some doubt over the practical feasibility of measuring this criterion in a comparable way (i.e., excluding any particular political or administrative circumstances) in different units. The university, in

this context, is interested in the amount of care provided, but only as an indicator of the time subtracted from research activity. In addition, the technical appropriateness of the medical care provided by university units is taken for granted (this is why we are able to speak of university teaching hospitals), and therefore this criterion is not an acceptable parameter for evaluation purposes.

In conclusion, the university is interested in the ability to deal with and the capacity to attract complex cases, because of the added value that these may confer on teaching and research activities. Having said that, even simple cases can provide valuable teaching opportunities for new medical graduates, even though straightforward situations are also readily found in other, non-university clinics within the regional health care system.

Furthermore, it has to be recognised that the ability to attract more complex cases does not depend only on the ability of the university professor, but also on the role, in the overall regional and corporate planning, that the unit is assigned, as well as on the simultaneous availability of similar units within the same hospital. Finally, the university might be interested in the ability to produce innovations, in particular with regard to patient care, that may not yet have been measured in terms of scientific production.

We might consider taking care activity volume as a measurable indicator, given that, from the perspective of the university units, it takes time and resources away from the research activity. The problem here is the difficulty providing a comparable indicator for different types of unit. Indeed, it is quite difficult to compare the number of admissions to a hospital unit with the number of outpatient services provided by a diagnostic unit, or the interventions of a psychiatric service with the projects handled by an epidemiological service. A possible solution, although difficult to implement, could be to translate each activity into costs or tariffs; but this would still not solve the problem, for example, of the anaesthesiology service or the psychiatric ward, whose interventions cannot be measured in terms of tariffs.

A further option is to consider the extent to which the resources assigned to the hospital match the volume of care provided by that hospital. A possible indicator in this case might be represented by the degree of coverage of the list of posts. This might prove to be a rather arbitrary criterion, since a clear organisation chart is often not available.

In any case, all these indicators of the volume of activity carried out are correlated with the amount of human resources available within each unit, which is already included in the evaluation criteria, and thus at risk of being considered twice.

Evaluation criteria and indicators of care delivery

Starting from the considerations set out in Box 4, five criteria for evaluating the medical care area can be identified (nature and complexity of the unit's management tasks, efficiency, complexity of cases dealt with, and ability to attract complex cases) as well as many related indicators.

The management activities of the unit can be evaluated:

1. with regard to their nature, using a quality indicator of the unit's complexity (for example as defined by the corporate deed: department/complex unit/departmental simple unit /simple unit);
2. with regard to their complexity, through a quantitative indicator of management tasks, for example the size of the staff to be managed.

Medical care activities could be evaluated applying three types of criterion and their relative quantitative indicators: efficiency, complexity of cases, and capacity to attract cases.

With regard to efficiency, indicators are needed that measure the efficiency of the various types of medical care activity, so as to compare them with other university healthcare units working in similar contexts. Therefore, the efficiency of a university cardiology unit is compared with that of another university cardiology unit, and not with that of a hospital cardiology unit; unfortunately, this is not always possible. In addition, since there exist at least three categories of primary care (inpatient, day-hospital and outpatient), it is necessary to choose a specific efficiency indicator for each activity within each unit, which should then be evaluated considering the distribution of this indicator regionally, in units from the same sector. Hence, the result is reported on a standardised scale for each category of activity and weighted by the importance of that category within the unit. This should make the indicator comparable among different units.

Once again, those structures whose activities (psychiatric field assistance, anaesthesiology services and epidemiology projects) do not fall into these categories, for which quantitative indicators are available, would be excluded. Efficiency evaluation, in these cases, could be done through a "*site visit*".

With regard to the complexity of cases, "*case-mix*" indicators could be used; however, in mixed hospital facilities, where university units often represent the only available care option for the given discipline, the *case-mix* is not determined by the university nature and specific competences of the unit, but rather by regional and corporate health programming decisions and patients' needs. In addition, comparable case-mix indicators are not available for all categories of care. That said, if we wish, in any case, to measure this criterion, indicators valid for each category of activity need to be identified: (*i*) case-mix indicator in the case of inpatient admissions, (*ii*) average number per day in the case of day-hospital admissions, and (*iii*) average number per day in the case of outpatient services. These indicators are measured in each unit, and then compared to the regional average of the units belonging to the same sector. Finally, the result is compared with a standardised scale of efficiency for each category of activity and assigned a category-based level of importance within the unit. This procedure should make it possible to compare different units. For the remaining activities, which do not fall into these three categories, the complexity of cases can be evaluated through a "*site visit*".

With regard to the capacity to attract cases, valid indicators can be used for each type of activity and unit; for example, it is possible to compute the proportion of non-standard users to overall users. These non-standard users can be inpatients in the hospital units, users of outpatient services in diagnostic units, and external projects in epidemiological units. Again, we encounter exceptions, in this case, unit whose usership is specified in regional or national health plans: in these situations, the capacity of attraction must again be carried out through a "*site visit*".

This evaluation system uses thus five synthetic indicators of care activities, which in turn are weighted by the number of university professors and researchers who make up the group and thus share the task of providing care.

Box 5
Evaluation of care at the "Polo San Luigi"

On the basis of data available at the hospital administration office, the following indicators are computed for each of the five elements to be evaluated:

1. *nature of the unit to be administered. This indicator is obtained from the corporate deed which classifies the units into four categories, to which importance is assigned in ascending order on the basis of the level of autonomy and responsibility: 0.15 for a simple unit, 0.30 for a departmental simple unit, 0.90 for a complex unit, 1.00 for a department administration;*
2. *the quantity of human resources to be administered is estimated on the basis of the amount of staff, relative to the previous year (employees or professional collaborators): this indicator also reflects the care volume;*
3. *efficiency is measured for inpatient admissions on the basis of the performance comparative index (obtained by comparison with units of the same discipline at regional level); for day-hospital admissions on the basis of the ratio between the average number of days per patient in each unit and the average numbers of days per patient recorded by similar units in the region; for specialist services through the number of examinations carried out in the unit out in relation to the staff-hours devoted to outpatient activity, compared to the average number recorded by other units in the region;*
4. *complexity of cases can be measured on the basis of the case-mix index for planned inpatient admissions, the average number per day for day-hospital admissions, and the average weight of invoices for outpatient services for the given specialty, compared with the values for other units of the same type in the region;*
5. *capacity to attract cases could be evaluated on the basis of the proportion of inpatient activity, outpatient services or project activity carried out in favour of users or clients not belonging to the group of users envisaged for the unit in regional and corporate programming.*

In each of the three cases (efficiency, case-mix, ca-

pacity of attraction), the indicators obtained for each type of activity (inpatient, day-hospital and outpatient) are standardised to a scale of 0 to 1, and are then weighted for the proportion of the total turnover generated, within the unit, by that particular type of activity (for example the amount of turnover of inpatient, day hospital or outpatient activities compared to the total turnover of the unit's activities) before being added together to give the standardised efficiency indicator.

In order to take account of the fact that each group may have a different number of university professors and researchers, the score of each synthetic care indicator (nature, importance, efficiency, case-mix, capacity of attraction) is related to the number of members of the group, for example by weighting the indicator with a value 1 if there is only one academic in the unit, 1.1 if there are two, 1.2 if there are three, and so on.

Ranking methods

Given the complexity of the evaluation process, the final ranking among structures can be produced within the framework of the *multi-criteria decision-making* (*MCDM*) models. These models are part of a broader class of operational research models, which deal specifically with decision-making issues in the presence of a specific number (limited) of criteria and alternatives. Basically, each method is a series of numeric techniques aimed at supporting the faculty in making a decision among a closed set of alternatives. The entire mechanism is based essentially on study of the decision-maker's utility function, underlying each of the MCDM methods.

Even though MCDM models have been criticised, mostly for their complexity, they are all widely implemented in various application sectors, from the allocation of resources in the health and environmental policy field, to the evaluation of surgical units based on a series of indicators (24). The Weighted Sum Model (WSM) is perhaps the most widespread and most used method. The Weighted Product Model (WPM) can be considered a modification of the WSM, some of whose theoretical limitations it was designed to overcome.

The Analytical Hierarchy Process (AHP) (25) is a further development of the WSM, and it is becoming increasingly popular. Other widespread methods are the *Elimination and Choice Translating Reality* (ELECTRE) (26) and the *Technique for Ordering Preference by Similarity to Ideal Solution* (TOPSIS) (27) methods.

Basically, MCDM models are built on a three-step procedure:

1. Understanding which element is useful for the evaluation purposes;
2. Associating numerical measures with each element;
3. Defining a method suitable for synthesising the results with a view to obtaining a possibly unique ranking of the alternatives: in our case, the SD sectors are the possible alternatives that the decision maker wants to rank.

All data are gathered in what is called the "decision matrix" **O** (28). This is nothing other than a matrix containing elements, or criteria (in the columns the C_1, C_2, C_3, …, C_n) and alternatives (in the rows O_1, O_2, O_3, …, O_m), in our case, the various groups or scientific sectors to be evaluated.

The measures available for all the alternatives and for each element, are indicated with o_{ij}, with $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$.

In addition, each criterion is weighted for its importance in the final evaluation with a value p_j, which totals 1 across all criteria.

	Criteria				
	C_1	C_2	C_3		C_n
Alternatives					
O_1	o_{11}	o_{12}	o_{13}	…	o_{1n}
O_2	o_{21}	o_{22}	o_{23}	…	o_{2n}
:	:	:	:	:	:
O_m	o_{m1}	o_{m2}	o_{m3}	…	o_{mn}
Weights	p_1	p_2	p_3	…	p_n

In the WSM, the best alternative corresponds to the highest score obtained by a linear combination of the elements o_{ij} and p_j

$$O^*_{WSM} = \max_i \sum_{j=1}^{n} o_{ij} p_j$$

The underlying assumption of an additive utility function, which underlies the WSM (in the sense that each measurement increases the overall utility for any given alternative by an amount proportional to the weight) is superseded by the WPM, where each alternative is compared with the others by multiplying a set of ratios elevated by the weight of the corresponding criterion (29). In particular, the first step is to compute the ratio of each pair of alternatives A_k e A_l for all criteria, and thus the product

$$R(O_K / O_L) = \prod_{j=1}^{n} (o_{Kj} / o_{Lj})^{pj}$$

whose maximum represents the best solution. A simpler alternative to the WPM is the so-called modified WPM or mWPM (30, 31), in which, instead of the ratios, each value is used directly as a term in the product

$$P(O_K) = \prod_{j=1}^{n} (o_{Kj})^{p_j}$$

The main drawback of the latter approach is that it loses the property of being an a-dimensional quantity, a property recognised in the full WPM.

The AHP is a linear approach similar to the WSM, in which the relative values are used instead of the crude ones used in the WSM.

Indeed, for each observed value o_{ij} the relative quantities are computed $q_{ij} = \frac{o_{ij}}{S_j}$ where $S_j = \sum_{j=1}^{n} o_{ij}$ The final AHP solution is thus

$$O^{AHP} = \max_i \sum_{j=1}^{n} q_{ij} p_j$$

The AHP method has been criticised because of a severe inconsistency, i.e., if an additional alternative is added to the set of previously considered alternatives, the final AHP solution can change, even if this additional alternative is not the optimal one. Belton and Gear (32) proposed the modified AHP as a solution to this problem. This approach consists of dividing each relative value q_{ij} by the maximum for each criterion. The final ranking is then obtained as in the standard AHP.

ELECTRE is based on the idea of a pairwise comparison of each alternative: in this case, the *outranking relationship* of two arbitrary alternatives O_i and O_j describes the situation in which where the decision maker is opting for one of the two. For each criterion, C_j, the difference between the crude and transformed values of each pair of alternatives for that given criterion is computed

$$g_i (O_j) - g_i (O_k)$$

Usually, before computing such differences, the values are normalised using the formula

$$x_{ij} = \frac{o_{ij}}{\sqrt{\sum_{i=1}^{m} o_{ij}^2}}$$

Then, the normalised decision matrix **X** is weighted, obtaining the weighted matrix

$$\mathbf{Y} = XP$$

Thus, the set of concordance C_{kl} is obtained for each pair O_k and O_l having dimensionality mk, with $l \geq 1$ and it is defined as the set of criteria for which O_k is preferred to O_l

$$C_{kl} = \left\{j, \text{ such that: } y_{kj} \geq y_{lj}\right\},$$

and the complementary set of discordance is defined oppositely as

$$D_{kl} = \left\{j, \text{ such that: } y_{kj} < y_{lj}\right\}$$

for all criteria $j = 1, 2, \ldots, n$.

The concordance index c_{kl}, which indicates the relative importance of the alternative O_k as compared to the O_l alternative, is thus the sum of weights p_j associated with the elements in the concordance set

$$c_{kl} = \sum_{i=C_{kl}} p_j$$

and it is always $0 \leq c_{kl} \leq 1$. The concordance indices, and the corresponding discordance indices, obtained similarly from the d_{kl} sets, are computed for each pair and put into two symmetrical matrices **C** and **D**. Each value is then compared with a threshold, or minimum acceptable performance, $f_{kl} = \mathrm{I}\ (c_{kl} \geq \underline{c})$, where the latter is determined by the relation

$$\underline{c} = \frac{1}{m(m-1)} \sum_{\substack{k=1 \\ k \neq l}}^{m} \sum_{\substack{l=1 \\ l \neq k}}^{m} c_{kl}$$

The same is then done for the discordance values. Finally, the aggregate dominance matrix **E**, having elements $e_{kl} = f_{kl} \times g_{kl}$, is used to identify the best alternative.

The TOPSIS method is based on the simple concept that the final ranking of SD sectors or groups should be structured in such a way that the highest rank has a "minimum distance" from the ideal situation, and a "maximum distance" from the non-ideal one. Although the distance concept may be interpreted in various ways, TOPSIS uses a Euclidean geometric interpretation. First, all values must be normalised to a-dimensional measures, obtaining a normalised matrix **R**, in which each element is computed as follows

$$r_{ij} = \frac{o_{ij}}{\sqrt{\sum_{k=1}^{m} o_{kj}^2}}.$$

The standardised and weighted matrix **V** is thus generated as

$$\mathbf{V} = \begin{bmatrix} p_1 r_{11} & p_2 r_{12} & p_3 r_{13} & \cdots & p_n r_{1n} \\ p_1 r_{21} & p_2 r_{22} & p_3 r_{23} & \cdots & p_n r_{2n} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ p_1 r_{m1} & p_2 r_{m2} & p_3 r_{m3} & \cdots & p_n r_{mn} \end{bmatrix}$$

Now, the ideal O^* and non ideal solution O^- must be defined as

$$O^* = \left\{ \left(\max_i v_{ij} \mid j \in J \right), \left(\min v_{ij} \mid j \in J' \right) \right\} = \left\{ v_1^*, v_2^*, v_3^*, \ldots, v_n^* \right\}$$

and

$$O^- = \left\{ \left(\min_i v_{ij} \mid j \in J \right), \left(\max v_{ij} \mid j \in J' \right) \right\} = \left\{ v_1^-, v_2^-, v_3^-, \ldots, v_n^- \right\}$$

where $i = 1, 2, \ldots, m$ and

$J = \left\{ j = 1, 2, \ldots, n \mid j \text{ is associated with "benefit criteria"} \right\}$

$J' = \left\{ j = 1, 2, \ldots, n \mid j \text{ is associated with "cost/loss criteria"} \right\}$

The "maximum" and "minimum" (J and J', respectively) situations are decided on the basis of two possible criteria: (*i*) a criterion defined prior to evaluation, during which the decision maker's absolute objectives are established; (*ii*) an empirical criterion, for which those situations correspond to the maximum and minimum registered for each criterion.

From the previous definition, it is deduced that the SD sector O^* is the best one, that is to say the ideal situation. In the same way, the SD sector O^- is the worst one, therefore the negative or non-ideal solution. The n-dimensional Euclidean distance is thus computed to measure the separation index of each SD sector from the ideal and non-ideal solution.

The formula to calculate the distance from the ideal solution is:

$$S_i^* = \sqrt{\sum_{j=1}^{n} \left(v_{ij} - v_j^* \right)^2}$$

The similarity of a specific SD sector O_i to the ideal solution O^*, is defined as:

$$C_i^* = \frac{S_i^-}{S_i^* + S_i^-}$$

where $0 \leq C_i^* \leq 1$ and $i = 1, 2, \ldots, m$. Clearly $C_i^* = 1$ if $O_i = O^*$, and $C_i^- = 0$ if $O_i = O^-$

The preference order can be ranked according to the C_i^* coefficients in descending order.

Among the MCDM models proposed in the scientific literature, the TOPSIS method perhaps shows greater elasticity in terms of its capacity not only to propose a ranking of units, but also to simulate a virtuous cycle of behaviours, indicating targets and priorities for the faculty. TOPSIS assumes that the behaviour of each SD sector or group has a tendency to increase or decrease the utility function of the "Polo San Luigi" according to a monotonic function, but this is usually not a heavy assumption, shared in any case by most of the methods discussed here. This assumption is however compensated for by the added flexibility of the TOPSIS approach: indeed, TOPSIS makes it possible to rank all sectors in relation to the ideal target with the highest and lowest utility and not only relative to the observed performances of the other competing sectors. The Euclidean distance criterion is thus used to evaluate the proximity of each SD sector to the ideal solution. The final ranking among SD sectors is obtained by comparing these relative distances.

The TOPSIS method therefore calculates Euclidean distances between each group and an ideal group, defined in two ways:

1. through an explicit and justified choice of a "target" value for each criterion, with the set of these target values constituting the "ideal solution";
2. through an empirical identification of the ideal solution as the combination of the maximum observed values for each criterion.

With regard to the present evaluation purposes, TOPSIS method requires the institution involved in the evaluation to choose explicitly the standards constituting the ideal group and the coefficients to be used to weight each criterion in terms of its importance.

Box 6
An application of the procedure at the "Polo San Luigi"

Data for the "Polo San Luigi" evaluation procedure are presented in Table 1. All the information was collected using a website, where each faculty member could enter his/her own data. Data on care activity were inserted thereafter by the "Polo San Luigi" Evaluation Committee and derived using the procedures described in Box 5. Weights for each of the criteria considered, discussed and formally approved by the "Polo San Luigi" faculty council, are also presented in Table 1. We applied, as described in the text, the TOPSIS method, which was chosen for its characteristics of flexibility and for its close reflection of the targets of the "Polo", which are not only to evaluate but also to stimulate excellence. In this first application, no specific autonomous targets were elicited by the faculty council of the "Polo" and thus the final ranking was produced with respect to the internal performances of each SD sector. An Excel file is available on the journal web site (http://www.bmscej.com/) as a complementary tool illustrating the ranking calculation steps.

The results are presented in Table 2 and they have been compared with both an additive utility approach (the AHP both in the standard and modified form) and the multiplicative utility scheme. The average distance between the maximum and the minimum ranking for each SD sector is 5.73 (range 2-12). The distance in ranking was much smaller comparing the TOPSIS with the AHP, on average 2.07 (range 0-6), due to the fact that both rely on Euclidean distances, than comparing the TOPSIS with the WSM, on average 4.27 (range 1-10).

Looking at the correspondence between the performances and the ranking, it can be noticed, that the SD sector SD-23, which the TOPSIS gives the highest ranking is overperforming in the teaching activity (the ratio of actual to designated hours is more than twice the average) and in the care activities, which together correspond to more than half of the total weights.

Discussion

In the case of the "Polo San Luigi, the elaboration of the evaluation system (clarification of the evaluation process criteria and data collection mechanism) was a long and laborious undertaking. Concrete application of the system then revealed some critical points most of which, it is hoped, future applications and the benefit of experience will make it possible to overcome.

The classifications elaborated using the TOPSIS method were, in general, comparable to those obtained applying the empirical methods used in the past for deciding how the budget should be shared: the groups in the first positions were still the same; at most, there was some switching of places at the top and bottom of the classification. The TOPSIS method, however, offers the advantage of allowing the specification of different weights and objectives in relation to the different stimulation and evaluation strategies, applied by the "Polo. As expected, the variable that most influenced the evaluation was the number of group members: this is obvious, since the distribution of the budget has to compensate for imbalances in the saturation of teaching, care and research loads (at comparable performance levels).

With regard to care activities, there are still numerous problems. Of the five selected evaluation criteria, only the size of the unit and the complexity of the staff to be managed can be measured with a degree of accuracy. The other three (efficiency, complexity of cases, capacity of attraction) are still measured quite imprecisely, due to the different mixes of activ-

Table 1. Values for each of the criteria explained in the text as observed in the evaluation procedure as applied at the "Polo San Luigi". Scientific sectors have been rendered anonymous and some data have been modified to ensure data masking.

	Scientific Activity			Teaching Activity						Care Activity			
	IF	COFIN	CIVR	Ratio of actual hours/ designated hours	ECT	Specialisation	PhD	Institutional Activity	Complexity	Personnel	Performance	Case-mix	Attraction
Weights	**30.0**	**3.0**	**2.0**	**29.0**	**1.5**	**1.5**	**1.5**	**1.5**	**6.0**	**10.5**	**4.5**	**4.5**	**4.5**
SD-1	54.01	1.0	1.0	0.98	0.0	0	0	0	0.71	0.71	0.43	0.43	0.57
SD-2	21.03	0.7	0.7	0.89	0.0	50	0	0	0.00	0.00	0.00	0.00	0.00
SD-3	9.74	0.7	0.7	0.84	6.0	50	0	0	1.20	0.53	0.41	1.01	1.21
SD-4	89.93	1.0	1.0	1.37	0.0	50	0	0	0.90	1.00	0.00	0.00	0.00
SD-5	29.05	1.0	1.0	1.29	0.0	50	0	20	0.82	0.55	0.53	0.56	0.56
SD-6	59.03	1.0	1.0	1.39	50.0	50	50	0	0.75	0.33	0.83	0.33	0.83
SD-7	58.41	1.0	1.0	0.73	0.0	50	0	0	0.00	0.00	0.00	0.00	0.00
SD-8	85.05	1.0	1.0	1.08	40.0	41	30	0	0.90	0.60	0.49	0.69	0.64
SD-9	15.78	0.7	0.7	1.04	12.0	50	50	0	1.00	1.11	1.11	1.11	1.11
SD-10	79.10	1.0	1.0	1.48	40.0	50	50	5	0.00	0.00	0.00	0.00	0.00
SD-11	83.91	1.0	1.0	0.95	0.0	50	0	0	0.00	0.00	0.00	0.00	0.00
SD-12	95.50	1.0	1.0	1.11	0.0	15	0	0	0.82	0.36	0.73	0.55	0.55
SD-13	87.20	1.0	1.0	1.27	0.0	50	0	0	0.90	0.60	0.00	0.00	0.00
SD-14	95.76	1.0	1.0	0.56	0.0	0	0	0	0.82	0.91	0.34	0.70	0.60
SD-15	31.12	0.7	0.7	0.50	0.0	24	0	0	0.00	0.00	0.00	0.00	0.00
SD-16	91.19	1.0	1.0	0.83	12.0	34	0	10	0.30	0.20	0.99	0.41	0.61
SD-17	52.06	0.7	0.7	0.46	0.0	50	0	0	1.09	0.97	0.48	0.71	1.03
SD-18	75.09	1.0	1.0	0.63	0.0	35	0	0	0.00	0.00	0.00	0.00	0.00
SD-19	53.45	1.0	1.0	0.59	0.0	18	0	0	0.00	0.00	0.00	0.00	0.00
SD-20	33.82	0.7	0.7	0.79	50.0	31	0	0	1.20	0.27	1.10	1.33	0.73
SD-21	71.98	1.0	1.0	0.96	0.0	0	0	0	0.82	0.18	0.72	0.58	0.49
SD-22	1.00	1.0	1.0	0.53	0.0	0	0	0	0.90	1.00	0.21	0.41	0.40
SD-23	27.36	1.0	1.0	2.98	0.0	0	0	0	0.91	0.91	0.51	0.38	0.18
SD-24	0.70	0.7	0.7	1.00	0.0	0	0	0	1.20	1.06	1.33	0.27	0.53
SD-25	0.70	0.7	0.7	1.00	0.0	0	0	0	1.20	1.33	0.00	0.00	0.00
SD-26	0.70	0.7	0.7	1.00	0.0	0	0	0	1.09	0.73	0.69	0.53	1.21
SD-27	0.70	0.7	0.7	1.00	0.0	0	0	0	1.33	0.53	0.00	0.00	0.00
SD-28	0.70	0.7	0.7	1.00	0.0	0	0	0	1.20	0.53	1.32	1.02	0.36
SD-29	0.70	0.7	0.7	1.00	0.0	0	0	0	1.09	0.97	0.62	1.00	0.50
SD-30	0.70	0.7	0.7	1.00	0.0	0	0	0	1.09	0.73	0.00	0.00	0.00

Abbreviations: IF = impact factor; COFIN = Co-Financing Scheme of the Italian Ministry of University and Research; CIVR = Italian University System Evaluation Committee; ECT = elective course teaching.

Table 2. Final scores and ranks.

Scientific disciplinary sectors	Final scores				Final ranks			
	WPM	AHP	AHP (Belton)	TOPSIS	WPM	AHP	AHP (Belton)	TOPSIS
SD-1	2.35	3.03	45.33	0.24	12	14	13	12
SD-2	0.89	1.57	20.26	0.04	22	30	29	29
SD-3	1.69	1.98	35.33	0.08	16	23	18	25
SD-4	2.59	4.56	59.91	0.57	7	1	2	2
SD-5	2.57	3.63	43.44	0.21	8	9	14	14
SD-6	3.50	4.18	54.57	0.40	2	5	7	7
SD-7	1.16	2.33	31.85	0.17	19	18	23	17
SD-8	3.83	4.43	60.59	0.46	1	4	1	6
SD-9	2.60	3.09	46.75	0.17	6	13	11	16
SD-10	2.06	4.45	48.78	0.49	15	3	9	5
SD-11	1.40	3.13	42.02	0.35	17	11	16	10
SD-12	3.05	3.94	59.04	0.50	3	7	3	3
SD-13	2.38	4.14	54.93	0.50	10	6	6	4
SD-14	2.49	3.74	57.07	0.39	9	8	5	9
SD-15	0.91	1.81	19.58	0.04	21	27	30	30
SD-16	2.71	3.45	52.99	0.40	4	10	8	8
SD-17	2.28	2.80	46.22	0.18	14	15	12	15
SD-18	1.20	2.59	35.65	0.24	18	16	17	13
SD-19	1.05	2.02	28.04	0.13	20	22	25	19
SD-20	2.38	2.59	42.67	0.13	11	17	15	18
SD-21	2.29	3.10	48.22	0.32	13	12	10	11
SD-22	0.59	1.61	25.96	0.04	27	29	26	28
SD-23	2.62	4.52	57.54	0.75	5	2	4	1
SD-24	0.70	2.13	34.62	0.11	24	20	19	20
SD-25	0.51	2.30	29.34	0.11	28	19	24	21
SD-26	0.69	1.88	32.70	0.09	25	24	21	24
SD-27	0.47	1.82	23.64	0.06	30	26	28	27
SD-28	0.68	1.78	32.30	0.09	26	28	22	23
SD-29	0.70	2.03	33.31	0.10	23	21	20	22
SD-30	0.48	1.88	24.08	0.06	29	24	27	26

Abbreviations: WPM = Weighted Product Model; AHP = Analytical Hierarchy Process; AHP (Belton) = Belton and Gear's modified AHP; TOPSIS = Technique for Ordering Preference by Similarity to Ideal Solution.

ities (inpatient, day hospital and outpatient/laboratory services) performed by each group, but also due to the poor availability of data at regional level for producing the indicators and to methodological inadequacies of the indicators used.

References

1. Aramaki T, Shimura T, Shimizu K, Miyashita T, Adachi K. [A report of questionnaire survey asking teaching staff concerning evaluation of faculty teaching from the students in Nippon Medical School]. J Nippon Med Sch, 2003; 70: 469-474. Published in Japanese.
2. Bardes CL, Hayes JG. Are the teachers teaching? Measuring the educational activities of clinical faculty. Acad Med 1995; 70: 111-114.
3. Colditz GA. The students' view of an innovative undergraduate medical course: the first year at the University of Newcastle, N.S.W. Med Educ 1980; 14: 320-325.
4. Faingold CL, Dunaway GA. Teaching pharmacology within a multidisciplinary organ system-based medical curriculum. Naunyn Schmiedebergs Arch Pharmacol 2002; 366: 18-25.
5. Herbert WN, McGaghie WC, Droegemueller W, Riddle MH, Maxwell KL. Student evaluation in obstetrics and gynecology: self-versus departmental assessment. Obstet Gynecol 1990; 76: 458-461.
6. Mash B, Marais D, Van Der Walt S, Van Deventer I, Steyn M, Labadarios D. Assessment of the quality of interaction in distance learning programmes utilizing the Internet or interactive television: perceptions of students and lecturers. Med Teach 2006; 28: e1-9.

7. McGee P. An evaluation of the lecturer-practitioner role in the independent health care sector. J Clin Nurs, 1998; 7: 251-256
8. Bezuidenhout J, Wasserman E, Mansvelt E et al. Clinical rotation in pathology: description of a case-based approach. J Clin Pathol, 2006; 59: 355-359
9. Glasson J, Chang E, Chenoweth L et al. Evaluation of a model of nursing care for older patients using participatory action research in an acute medical ward. J Clin Nurs 2006; 15: 588-598.
10. Levine SA, Caruso LB, Vanderschmidt H, Silliman RA, Barry PP. Faculty development in geriatrics for clinician educators: a unique model for skills acquisition and academic achievement. J Am Geriatr Soc 2005; 53: 516-521.
11. Stineman M, Kennedy B. Response to "Metrics of rehabilitation research capacity": within and beyond our borders. Am J Phys Med Rehabil 2005; 84: 1020-1033.
12. Folse JR, DaRosa DA, Wood M. An evaluation of surgery departments in community-based medical schools. Arch Surg 1991; 126: 1122-1127.
13. Heagarty MC, Boehringer JR, Lavigne PA, Brooks EG, Evans ME. An evaluation of the activities of nurses and pediatricians in a university outpatient department. J Pediatr 1973; 83: 875-879.
14. Houston TK, Ferenchick GS, Clark JM et al. Faculty development needs. J Gen Intern Med 2004; 19: 375-379.
15. Kelley WN, Stross JK. Faculty tracks and academic success. Ann Intern Med 1992; 116: 654-659.
16. Mahler S, Neumann L. Faculty development in medical instruction: the cognitive dimension. Isr J Med Sci,1987; 23: 1247-1251.
17. Maker VK, Curtis KD, Donnelly MB. Faculty evaluations: diagnostic and therapeutic. Curr Surg 2004; 61: 597-601.
18. McLeod PJ, Steinert Y, Nasmith L, Conochie L. Faculty development in Canadian medical schools: a 10-year update. CMAJ 1997; 156: 1419-1423.
19. Ohara S, Kawaguchi Y, Shinada K, Sasaki Y. Evaluation of school-based dental health activities including fluoride mouth-rinsing in Hiraizumi, Japan. J Med Dent Sci 2000; 47: 133-141.
20. Steinert Y, Cruess S, Cruess R, Snell L. Faculty development for teaching and evaluating professionalism: from programme design to curriculum change. Med Educ 2005; 39: 127-136.
21. Tisonova J, Szalayová A, Kriska M, Gibala P, Foltán V. Factors influencing the spontaneous reporting of adverse drug reactions - the experience of the Slovak Republic. Pharmacoepidemiol Drug Saf 2004; 13: 333-337.
22. Yonemitsu M, Kawaguchi Y, Ohara S et al. [Evaluation of school dental health activities in Hiraizumi primary school, Iwate prefecture]. Kokubyo Gakkai Zasshi, 1992; 59: 562-570. Published in Japanese.
23. Maker VK, Lewis MJ, Donnelly MB. Ongoing faculty evaluations: developmental gain or just more pain? Curr Surg 2006; 63: 80-84.
24. Toniati W, Gregori D. Calcolo dell'offerta economicamente più vantaggiosa. Milan: IPSOA 2002.
25. Saaty TL. Fundamentals of decision making and priority theory with the AHP. Pittsburgh, PA, USA: RWS Publications 1994.
26. Benayoun R, Roy B, Sussman N. Manual de reference du programme Electre. Note de synthèse et formation. Vol. 25. 1966, Paris, France: Direction Scientifque SEMA 1966.
27. Hwang CL, Yoon K. Multiple Attribute Decision Making: methods and applications. 1981, New York, NY, USA: Springer-Verlag.
28. Zimmermann HJ Fuzzy Set Theory And Its Applications. Third ed. 1996, Boston, MA, USA: Kluwer Academic Publisher.
29. Bridgman PW. Dimensional Analysis. 1922, New Haven, CT, USA: Yale University Press.
30. Barzilai J, Lootsma FA. Power relations and group aggregation in the multiplicative AHP and SMART. in Third International Symposium on the AHP. 1994. Washington DC: George Washington University.
31. Lootsma FA. Multi-Criteria Decision Analysis via Ratio and Difference Judgment. Applied Optimization Series. 1999, Dordrecht, The Netherlands: Kluwer Academic Publisher.
32. Belton V, Gear T. On a short-coming of Saaty's method of analytic hierarchies. Omega 1983; 11: 228-230.