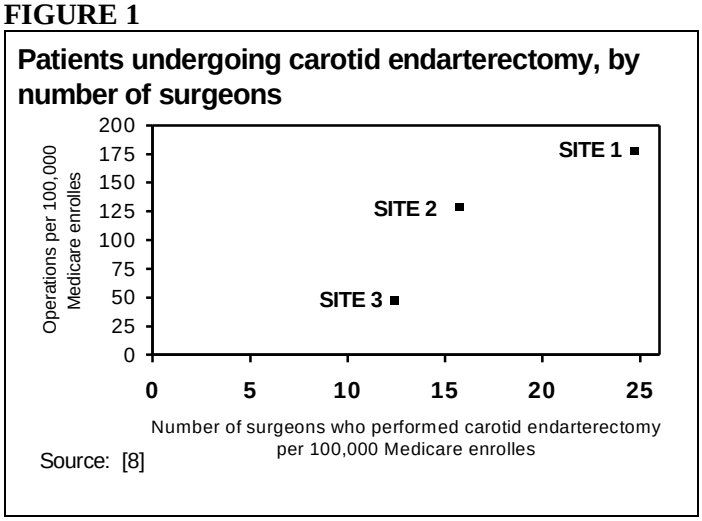
OUTCOMES RESEARCH AND APPROPRIATENESS OF MEDICAL TECHNOLOGIES[*]

Pablo Lázaro y de Mercado[**], MD, MBA, Ph.D.
Health Services Research Unit
Instituto de Salud Carlos III

Abstract: As health expenditures continue to consume increasingly large proportions of national budgets, ways must be found to assure that resources spent on health are used for effective services. Since only a small proportion of medical decisions are based on scientific evidence about their outcomes, it is not surprising that wide variations exist in clinical practice. In recent years, questions have been raised about how medical decisions are made and the proportion of medical procedures that are performed for appropriate reasons. One method developed to quantify the amount of inappropriate use is the so-called "RAND appropriateness method," which is based on the scientific evidence and the collective judgment of an expert panel. Measured by this method, a number of procedures have been shown to have high rates of inappropriate or uncertain use. The challenge is to find ways to eliminate ineffective services to free resources for those that have been proven effective. Further research is needed to improve the method and to find acceptable ways its findings can be used to promote effective care.

Health services face challenges such as increasingly complex services, limited resources, and rapid innovation and diffusion of medical technologies [1,2]. Added to this is the fact that in the last 25 years spending on health has grown twice as fast as wealth in the industrialized countries [3]. Although there is no single "correct" proportion of wealth that health care should consume, it is clear that this growth must be stabilized at some point. Ideally, the required savings would be found by correcting administrative and operational inefficiencies, but most of the excess growth in health expenditures is due to increases in the "volume and intensity of services" [4]. To curtail the volume of services without negative effects on the health status of the population, it will be necessary to find ways to assure that our health resources are used for effective services, that is, those that have demonstrated value [5]. The challenge is not to deny care, but to find a way to do it that not only avoids harming quality, but enhances it. Our problem is that we know very little about the determinants and outcomes of the application of health technology [6]. It has been calculated that only around 15% of medical decisions are based on scientific evidence about their outcomes [7]. Thus, it is not surprising that such wide variations exist in clinical practice. As an example of clinical practice variations, figure 1, which plots rates of carotid endarterectomy as a function of the number of surgeons in three sites of the United States (US), shows wide variations in the use of this technology and a positive association between number of surgeons and number of surgical procedures performed [8]. In this examples one must ask: What is the most reasonable rate of use? Are some persons receiving unnecessary surgery, and/or are others not receiving surgery they need?

FIGURE 1



Variations in clinical practice occur in every country. In Spain, for example, wide variations exist in the rate of mammography use for women aged 40 to 70 by geographic region, ranging from 12% in the region with the lowest use, to 74% in the region with the highest [9,10]. Clinical practice variations may have a dramatic effect on the use of medical technology. For example, physicians who have an x-ray device in their office request a chest x-ray for 46% of their patients as compared to only 11% when the physician must refer the patient to a radiologist

[*] Lázaro P. Outcomes research and appropriateness of medical technologies. In: *The place of outcomes research in health technology assessment*. Barcelona: International Society for Technology Assessment in Health Care;1997:147-56.

[**] Current address: Pablo Lázaro. Técnicas Avanzadas de Investigación en Servicios de Salud (TAISS). C/Cambrils 41-2, 28034 Madrid. Spain. E-mail: plazaro@taiss.com

[11]. Again the questions arises: What is the right proportion of procedures to be requested, 46% or 11%? Are there patients underserved in the radiologist-referral group of physicians, or is there an excess of x-ray exams in the self-referral group?

Wide variations among countries have also been observed [12,13]. For example, there is enormous variation in the distribution of "big ticket" medical technologies in OECD countries, whether measured in terms of population unit, health care expenditures, or wealth [13]. In a case study of this phenomenon in Spain, it was found that the country ranked among the top OECD countries in number of extracorporeal shock wave lithotripter (ESWL) units, whereas it ranked among the last countries in radiation therapy units. This study showed that the payment incentives designed by public sector health authorities in Spain permitted large profits for private providers of ESWL treatment, but no profit for private providers of radiation therapy. One of the consequences of this distorted system of financial incentives is that Spain has an abundance of lithotripsy units and a scarcity of megavoltage units in comparison to other OECD countries. Furthermore, these technologies are distributed unevenly throughout Spain: the wealthier regions have significantly more lithotripters per population unit than the poorer ones. And finally, 72% of the lithotripsy units in Spain are installed in the private sector, versus only 16% of the radiation therapy facilities [14]. One of the conclusions of this study was that the inequitable distribution of the two technologies could be explained by inadequately designed financial incentives for providers. In the first place, there are unreasonable differences between price and cost, that is, in profit, for private providers of these services; and second, there is no relationship between payment for the procedure and its appropriate use. As in the previous examples, does this mean that cancer patients in Spain do not receive appropriate radiation therapy care in comparison with the OECD population? Or are patients in OECD countries receiving unnecessary treatment? Is the situation reversed in the case of ESWL?

TABLE 1

Ratings of appropriateness of four procedures, United States (%)

Procedure	Appropriate	Uncertain	Inappropriate
Coronary angiography	74	9	17
Carotid endarterectomy	35	32	32
Upper GI endoscopy	72	11	17
Hysterectomy	58	25	16

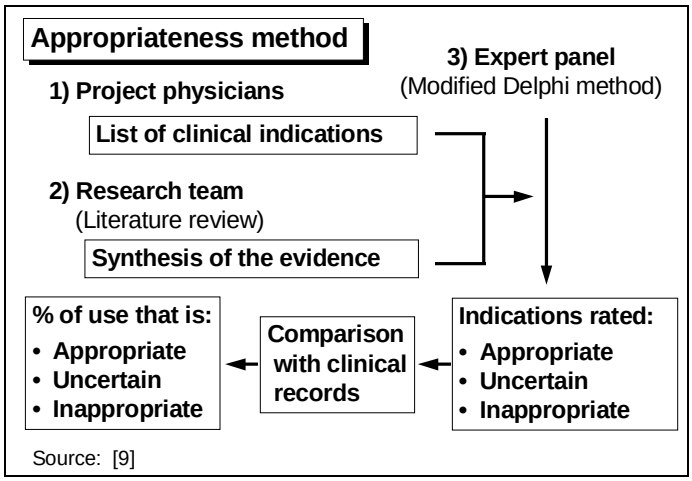
Sources: [15, 16]

In an attempt to answer these kinds of questions, researchers from RAND and UCLA have developed what is known as the "RAND appropriateness method" (RAM). Using this method, a high proportion of medical procedures in the US. were judged to be inappropriate or uncertain (Table 1). In the case of coronary angiography, 17% of the procedures were done for inappropriate reasons and 9% were uncertain. For carotid endarterectomy, 32% were inappropriate and another 32% were uncertain, while 17% of upper gastrointestinal endoscopies were inappropriate and 11% were uncertain [15]. Another study found that 16% of hysterectomies performed in the US were done for reasons judged to be clinically inappropriate and 25% were uncertain [16].

These rates were determined by applying the RAM, which is based on a literature review and the collective judgment of an expert panel [9,17]. Figure 2 shows an overview of the different steps used in carrying out this method. First, several physician-specialists in the field under study develop a list of all possible clinical indications for the procedure, categorized in terms of patients' symptoms, the results of diagnostic tests, and their previous medical history. For example, a Spanish study designed to study the appropriateness of surgery for benign prostatic hyperplasia divided the indications into seven main chapters: Acute Urine Retention, Chronic Retention, Hematuria, Urinary Infection, Bladder Stones, Diverticula, and Symptoms in the absence of the preceding conditions [18]. The chapters were then subdivided into additional sections. An example of a specific clinical indication used in this study is: a patient with acute urine retention, suffering a repeated episode, with retention of more than 500 cc of urine, with a life expectancy of 1-4 years, and moderate symptoms. Another example, taken from a US study of the appropriate use of coronary artery bypass graft surgery (CABG), is: a patient with chronic stable angina, class I/II, treated with maximal medical therapy, three-vessel disease, ejection fraction >35%, candidate for percutaneous transluminal coronary angioplasty (PTCA), and low risk" [19]. It is essential that the indications be both comprehensive enough to permit classification of all patients and mutually exclusive, so that no patient can be classified in more than one category. For the procedures studied so far, the number of indications typically exceeds 1000.

After the list of indications is developed, the research team makes a systematic, critical review of the literature and produces a synthesis of the scientific evidence to date. The articles are usually obtained through MEDLINE searches for a specific period and manual searches of bibliographic citations. The articles identified are sorted according to the strength of the evidence, with most priority given to the results of large, well-conducted, randomized controlled trials. The information is organized by outcomes, risks, utilization, and other relevant aspects of the procedure. The purpose of the synthesis of the scientific evidence is to provide panel members with the most up-to-date knowledge available to help them in the rating process.

FIGURE 2



Once the list of indications and the synthesis of the evidence are completed, both documents are sent to an expert panel, starting the third major step in the method. Expert panels generally consist of nine members, including both specialists and general practitioners. Some specialists practice the procedure, and some do not (e.g., panels on the appropriateness of cardiology procedures would include both interventionist cardiologists and general cardiologists). The panelists are selected based on their clinical experience and recognized prestige in the scientific community. Depending on the particular country, it may also be desirable to have representation from different geographic regions, types of hospitals, or other characteristics. Nominations are often solicited from professional groups.

Using a modified Delphi process, the indications are rated in two rounds. In the first round, the report synthesizing the available evidence and the list of indications are sent to the panel members, together with instructions on the rating process. The panel members rate each indication on a scale of 1 to 9 according to how appropriate they consider the performance of the procedure to be for that particular indication. All ratings are confidential and the panel members are unknown to one another. A rating of 1 means the procedure is extremely inappropriate for the indication, a 9 means that it is extremely appropriate, and ratings between 4 and 6 indicate uncertainty as to whether the procedure is appropriate or not. Each panel member receives a written definition of "appropriate" for the purposes of this study, that is, "The expected health benefit, such as increased life expectancy, relief of pain, or improved quality of life, exceeds the expected negative consequences, such as mortality, morbidity, or anxiety, by a sufficiently wide margin that the procedure is worth doing, exclusive of cost." In the second Delphi round, panelists meet for one or two days to discuss and re-rate the indications. Each panelist receives a document showing his or her own ratings, the group median, and a summary of all responses for each indication. The panel moderator points out areas of confusion or disagreement for discussion. The panelists are given the opportunity to modify the structure of the list of indications by expanding, collapsing or adding new categories, if desired. At the end of the session, the panelists again rate each indication in the final list. It should be emphasized that the entire process is designed to identify agreement, but not to force panel members to reach a consensus.

After the second Delphi round, the indications are classified as "appropriate," "inappropriate," or "uncertain" depending on the median rating and level of agreement with which they are rated. An indication is considered to be rated "with agreement" (for a 9-member panel) when no more than two panelists rate it outside the 3-point region containing the median. Disagreement occurs when three or more panelists' ratings are in the 1-3 region, and three or more are in the 7-9 region. Indications for which there is neither agreement nor disagreement are considered to be "indeterminate." All indications which are rated *without disagreement* are classified as appropriate if the median score is 7-9, and inappropriate if the median score is 1-3. Indications with a median of 4-6, as well as all those rated *with disagreement*, regardless of the median, are classified as uncertain.

One question that arises is what effect different panels and different cultural contexts have on the results of this process. An interesting study in this respect compared the results of applying two sets of appropriateness criteria--one from a US panel and one from the United Kingdom (UK)--to the same set of patients receiving CABG in the UK. Because the rate of use of this procedure is considerably lower in the UK, it was hypothesized that few of the procedures would be judged inappropriate. Surprisingly, it was found that both groups judged a substantial proportion of bypass surgeries to be inappropriate--7% by the US criteria and 16% by the UK criteria, or

uncertain--26% vs. 27% [20]. In analyzing why the two panels rated the same indications differently, it was found that most of the differences occurred when there was little scientific evidence about the outcomes of the procedure.

In addition, these figures show that just reducing the number of procedures performed will not necessarily reduce the rate of inappropriate use. Thus, another question is why, in situations of severely restricted resources, fixed budgets, and salaried physicians, as in the case of the UK, procedures continue to be used for less than appropriate reasons. One possible explanation could be the absence of guidelines or standards to guide physicians in the appropriate use of different procedures. In the absence of such standards, it is not surprising that inappropriate use can occur side by side with rationing and underuse. The comparison between the US and the UK suggests that budgetary limitations or different types of economic incentives for providers may decrease the number of procedures, but not necessarily the proportion of them that are inappropriate.

The effects of economic incentives on consumers of health care have also been studied, but the results have been rather disappointing. For example, in the large Health Insurance Experiment carried out in the US, researchers hypothesized that if people had to pay more for health care, they would selectively reduce their demand for care for conditions that medical science can do little about. They found instead that demand was reduced about equally for both medically effective and ineffective services [21]. These examples suggest that economic incentives alone, whether for providers or consumers, will not be enough to improve the appropriateness of care.

TABLE 2

Appropriateness of Coronary Artery Bypass Surgery in the U.S.

	A western state (1979-82)	New York state (1990)
Appropriate (%)	**56**	**91**
Uncertain (%)	**30**	**7**
Inappropriate (%)	**14**	**2.4**

Sources: [22, 19]

Many factors, including policy interventions, can affect the proportion of inappropriateness. For example, it is interesting to compare the results of two studies of the appropriateness of CABG in different parts of the US (Table 2). The earlier study, which included patients from three hospitals in a western state in the early 80s, found 14% of the CABG procedures reviewed to be inappropriate [22], versus only 2.4% of those carried out in a 1990 study in New York State [19]. A much higher proportion of procedures was also rated uncertain in the western state study: 30% as compared to 7% in New York. There are several possible explanations for these differences. First, it is probable that appropriateness ratings in the latter study changed in response to new scientific evidence. Second, practice patterns have changed in response to the rapid rise of PTCA between the two periods of study. Third, it may be that the selection of patients for CABG is different in New York than in other states. This hypothesis is supported by the fact that the New York Department of Health limits the number of cardiac surgical centers, sets high credentialing standards, and reports risk-adjusted mortality data by hospital and surgeon [19]. These measures provide strong incentives for each hospital to monitor its own performance, and may help explain why New York had half the rate of CABG procedures as the national average, and a very low proportion of inappropriate procedures. In sum, a policy intervention involving information, explicit standards, and incentives can reduce the number of procedures, mainly by eliminating inappropriate ones, and consequently would reduce costs and at the same time increase the quality and efficiency of the system.

Combining data from different studies, it has been estimated that roughly one-third of health care expenditures goes to services that are of little or no benefit [23]. Thus, the conclusion is not necessarily that the provision of health services must be rationed, but that the selective elimination of ineffective services would free resources to care for those who need effective diagnostic or therapeutic technologies. This kind of policy would not only increase the efficiency and quality of the health system, but would also make it more equitable by avoiding the restriction of effective services. Rationing, whether it is explicitly imposed by requirements of co-payments or deductibles as in the US, or implicitly imposed by the need to join long waiting lists for services as frequently happens in Europe, will have a disproportionate effect on the poor, the elderly, and the chronically ill [21].

Achieving such improvements will require sustained funding for research. Appropriateness projects are expensive and lengthy. But the investment is well worth the effort if we look at the potential savings. In Spain, for example, a research project on the appropriateness of CABG and PTCA takes about 18 months and costs nearly half a

million dollars. However, if only 10% of the procedures are found to be inappropriate and are selectively eliminated, about 13 million health care dollars can be saved per year [24].

In sum, it has been seen that wide variations exist in clinical practice. It is estimated that a high proportion of health services are performed for inappropriate reasons. Bureaucratic, administrative, or economic solutions to rising costs may limit the quantity of health care provided, but will not necessarily improve the appropriateness and quality of care. The selective elimination of inappropriate care would free resources to deliver effective care to those who need it. But if the appropriateness method is to be used, further research needs to be done to improve the method and to find acceptable ways in which its findings can be used to promote effective care and reduce that which is inappropriate. The development of such standards could provide physicians and policy makers with a flexible tool that could be used to reduce the number of procedures performed for inappropriate reasons. Such an effort will require the development of both clinical standards and a way to implement them that is acceptable to physicians. Cooperation among physicians, government, and the public will be essential if we are to achieve the goal of a more efficient and equitable use of our health resources.

REFERENCES

1. Lázaro P, Pozo F, Ricoy JR. Una estrategia de investigación en el sistema nacional de salud: II. Investigación en servicios de salud. Med Clin (Barc) 1995;104:67-76.
2. Lázaro P. Evaluación de Tecnología Médica. Valencia: M/C/Q ediciones; 1994.
3. OECD HEALTH DATA. A software package for the international comparison of health care systems. Version 1.5. Paris, France, OECD, 1993.
4. Eddy DM. Broadening the responsibilities of practitioners. The team approach. JAMA 1993;269:1849-1855.
5. Lázaro P, Azcona B. Clinical practice, ethics, and economics: the physician at the crossroads. Health Policy 1996;37:185-198.
6. Pozo F, Ricoy JR, Lázaro P. Una estrategia de investigación en el sistema nacional de salud: I. La epidemiología clinica. Med Clin (Barc) 1994;102:664-669.
7. Black N. Research, audit, and education. BMJ 1992;304:698-700.
8. Leape LL, Park RE, Solomon DH, Chassin MR, Kosecoff J, Brook RH. Relation between surgeons' practice volumes and geographic variation in the rate of carotid endarterectomy. NEJM 1989;321:653-657.
9. Lázaro P, Fitch K. From universalism to selectivity: is “appropriateness” the answer? Health Policy 1996;36:261-272.
10. Luengo S, Lázaro P, Madero R, Alvira F, Fitch K, Azcona B, Pérez JM, Caballero P. Equity in the access to mammography in Spain. Soc Sci Med 1996;43:1263-1271.
11. Hillman BJ, Joseph CA, Mabry MR, Sunshine JH, Kennedy SD, Noether M. Frequency and costs of diagnostic imaging in office practice--a comparison of self-referring and radiologist-referring physicians. NEJM 1990;323:1604-1608.
12. Lázaro P. Evaluación de Servicios Sanitarios: La Alta Tecnología Médica en España. Madrid: Fondo de Investigación Sanitaria; 1990.
13. Lázaro P, Fitch K. The distribution of "big ticket" medical technology in OECD countries. Int J Tech Ass Health Care 1995;11:552-570.
14. Lázaro P, Fitch K. Economic incentives and the distribution of extracorporeal shock wave lithotripters and linear accelerators in Spain. Int J Tech Ass Health Care 1996;12:735-744.
15. Chassin MR, Kosecoff J, Park RE, Winslow CM, Kahn KL, Merrick NJ, Keesey J, Fink A, Solomon DH, Brook RH. Does inappropriate use explain geographic variationsin the use of health care services? A study of three procedures. JAMA 1987;258:2533-2537.
16. Bernstein SJ, McGlynn EA, Siu AL, Roth CP, Sherwood MJ, Keesey JW, Kosecoff J, Hicks NR, Brook RH. The appropriateness of hysterectomy. JAMA 1993;269:2398-2402.
17. Brook RH, Chassin MR, Fink A, Solomon DH, Kosecoff J, Park, RE. A method for the detailed assessment of the appropriateness of medical technologies. Int J Tech Ass Health Care 1986;53-63.
18. Berra A, Lázaro P, Fitch K, Martin A, Calahorra L. Appropriate indications for surgery of benign prostatic hyperplasia. ISTAHC 11th Annual Meeting. Stockholm, Sweden, 4-7 June 1995 (Abstract No. 16 and oral presentation).
19. Leape LL, Hilborne LH, Park RE, Bernstein SJ, Kamberg CJ, Sherwood M, Brook RH. The appropriateness of use of coronary artery bypass graft surgery in New York State. JAMA 1993;269:753-760.

20. Bernstein SJ, Kosecoff J, Gray D, Hampton JR, Brook RH. The appropriateness of the use of cardiovascular procedures: British versus US. perspectives. Int J Tech Ass Health Care 1993;9:3-10.

21. Lohr KN, Brook RH, Kamberg CF, Goldberg GA, Leibowitz A, Keesey J, Reboussin D, Newhouse JP. Use of medical care in the Rand Health Insurance Experiment. Diagnosis- and service-specific analyses in a randomized controlled trial. Medical Care 1986;24(9 Suppl):S1-S7.

22. Winslow CM, Kosecoff J, Chassin M, Kanouse DE, Brook RH. The appropriateness of performing coronary artery bypass surgery. JAMA 1988;260:505-509.

23. Brook RH, Lohr KN. Will we need to ration effective health care? Issues in Science and Technology 1986;3:68-77.

24. Lázaro P. Angioplastia coronaria y cirugía coronaria: algunas consideraciones socio-económicas. Rev Esp Cardiol 1993;46 (supl. 3):1-14.