ELSEVIER

Health Policy 35 (1996) 247–265

HEALTH policy

The ambiguous relationship between practice variation and appropriateness of care: an agenda for further research

Anton F. Casparie

Institute for Health Policy and Management, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Received 6 May 1995; revision received 7 August 1995; accepted 14 August 1995

Abstract

The analysis of variation in the use of health care services, and particularly of practice variation, has been the subject of renewed interest because of the view that the inappropriate use of procedures could be a major cause of these differences. In this article, recent literature is reviewed and the results of personal research are described on both the variation in care provision and on appropriateness assessment. In the few studies which have focussed on both subjects no evidence has yet been found to suggest that practice variation is to be explained by differences in appropriateness rates. However, there are still many methodological pitfalls in both variation analyses (statistical problems) and appropriateness assessment (reliability of the judgement), implying that this conclusion is far from definitive. More research should therefore be conducted on methodological questions of variation analysis and appropriateness assessment. Furthermore in variation analysis the relative contribution of all potential determinants has to be studied on the various levels of care provision. Finally, to study the relationship between practice variation and appropriateness of care, the clinical problem and not the procedure should be the starting point.

Keywords: Small area variation; Practice variation; Appropriateness of care; Quality of care

* Corresponding author. Tel.: +31 10 4088525; fax: +31 10 4526086.

SSDI 0168-8510(95)00787-3

1. Introduction

There have always been differences in health care between geographic areas, institutions and individual providers, both in the structure, process and outcome of care. Differences in structure include, for example, the number of beds available. Process differences denote variations in the use of the health care facilities available. Differences in outcome are related to differences in mortality between regions or socio-economic classes, or differences in mortality rates between hospitals.

In this article, the differences in health care to be examined will be restricted to the process of care: the differences in the use of health care services, and particularly the patterns of practice. These differences are the result of a complex interaction between the demand for, and supply of, care. One possible reason for these differences could be the variation in style of practice between various medical specialists, due to divergent indications of diagnostic and therapeutic procedures with the possible consequence of differences in the quality of care.

The question addressed in this advocacy paper is: is there evidence that differences in the use of health care services can be explained by differences in the quality of care, especially the appropriate use of diagnostic and therapeutic procedures? With this purpose, first an overview is given of the recent literature on variation in the use of health care services, and on measurement on the appropriate use of diagnostic and therapeutic procedures. Next, the existing research on the potential relationship between practice variation and appropriateness of care will be discussed. The methodological pitfalls of current research on these topics will be described and it will be shown that, at present, no definitive conclusion can be drawn about such a relationship. Therefore, for each item an agenda for further research is proposed.

Computerized searches were conducted of the relevant English-language literature published between 1988 and 1994, using the Medline databases. Furthermore, every issue of the 1993 and 1994 volumes of the *British Medical Journal*, *Health Services Research*, *Journal of the American Medical Association*, *Journal of Quality Improvement*, *International Journal for Quality in Health Care*, *New England Journal of Medicine*, and *Medical Care and Quality in Health Care* were perused for articles related to this subject area. The emphasis of this article will therefore be placed upon the results published in recent literature. Data from the literature will be further illustrated with some results of our own research.

2. Differences in the use of health care services

2.1. Recent evidence

The existence of differences in the use of health care services has been a topic of study for many years [1]. However, over the past few years renewed interest has been shown for this phenomenon from the point of view of cost-containment and efficiency of care.

In the literature, terms such as geographic variation, small area variation, variation between institutions and practice variation have been used to describe

these differences. However, a distinction based upon level is also necessary as the reasons or determinants for the variation can differ from level to level. Geographic variation denotes differences between countries or large areas; it is conceivable that, on this level, differences between health care systems, or in the morbidity of the population, will come to play a prominent role. On the other hand, differences on the institutional level, or between individual providers, could more probably be attributed to variations in attitudes or opinions prevalent between individual physicians. In most reports, the distinction between the various geographic levels is not mentioned explicitly when conclusions have been drawn. Not all authors seem to be aware of the importance of this distinction.

A considerable amount of literature is available on small area variation, in particular that published by the group lead by Wennberg. To give some idea of the degree of differences on the various geographic levels, as well as the level of recent interest shown in this subject, the three studies published by this group in 1993 and 1994 will be remarked upon [2–4] together with another recent article in this area [5]. Some other recent studies will be discussed in the section on the determinants of small area variation and practice variation.

One study that concerns relatively small areas found a twofold difference in expenditure on physicians' services to Medicare patients in 317 U.S. metropolitan areas, when adjustments were made for age and sex and, with regard to inpatient care, also adjusted for patient-mix [2]. There was no substitution between inpatient and outpatient care. In another study on larger regions, physicians in Florida turned out to use one and a half times more resources per admission for inpatient treatment than their colleagues in Oregon, when adjusted for the physicians' case-mix [3]. In a third study by the Wennberg group, patients in Boston who were initially hospitalized with acute myocardial infarction, a stroke, gastrointestinal bleeding, hip fracture or cancer surgery had a 1.64 higher chance of readmission in the next 3 years than comparable patients in New Haven [4]. Research by Poses et al. [5] discovered an almost twofold difference in the prescription of antibiotics between two HMOs, one in Pennsylvania (32.4%) and the other in Nebraska (72.9%); these are both very small areas.

In the Netherlands, interest in small area variation was aroused after the publication of '*Small Area Variation in the Use of Hospital Facilities*', a report compiled in 1989 by a research committee of the Royal Dutch Medical Association [6]. In this report the areas studied contain about 500 000 inhabitants. We studied the differences in patterns of practice found in the specialist partnerships of six medical specializations in all of the 10 hospitals situated in one health insurance region of the Netherlands. More detailed results of this study, which was directed towards variation on the group physician level, have been published elsewhere [7,8]. Wide differences were found in the number of admissions, patient days and the in- and outpatient procedures employed by the physicians of a single specialization. These differences diminished somewhat when adjusted for age, sex and number of partnership members (as a substitute for morbidity of the treated patients), but nevertheless remained substantial. The smallest difference found was a factor of 1.1 in patient days within the specialization of pulmonology, with the largest being

among cardiologists with a factor of 6.8 for hospital admissions. There was no indication of substitution between in- and outpatient care as a reason for these differences. Furthermore, we found no obvious distinction in differences between four specific procedures with a supposedly clear, and five with a less-defined, indication. This finding contrasts with the opinion that the major causes of practice variation are related to a lack of agreement on the optimum treatment to be offered.

In a more recent study, we looked for variations in the diagnosis and treatment of patients with benign prostate hyperplasia by 12 urologist partnerships in 13 hospitals scattered around the Netherlands. In this research, the level of investigation is more the small area in which differences in morbidity, as well as practice variation, could play a role. There turned out to be large differences in performance rates of diagnostic procedures between these 12 urologist groups, especially for ultrasonography. Preliminary results show that ultrasonography of the kidney was used by the partnership with the lowest rating on just 8% in their patient population while the partnership with the highest rating performed this on 80% of their patients. For prostate ultrasonography these figures were 17 and 90%, respectively. Even for cystoscopy the rates varied substantially; 26 against 68%.

In addition, the initial therapeutic choice in these patient populations also differed between the various specialist groups: transurethral resection ranged from 20 to 48%, and initial drug treatment (5-α reductase inhibitor or α-blocker) from zero to 36%.

2.2. Methodological pitfalls

Diehr and coworkers have extensively researched the reliability of data from small area analysis [9–12]. They used a computer programme to simulate the distribution of commonly used descriptive statistics to test their null hypothesis that all the small areas have similar rates. In their study the chi-square statistics had the highest power [11,12]. The expected variability was surprisingly large, especially for procedures with low incidence, for smaller populations, where there is variability among the populations of the areas, and where more than one procedure per inhabitant is possible (e.g. readmissions). Caution should therefore be taken with studies of low incidence events and smaller populations. However, even if there is a statistically significant variation, the size or importance of the differences among the small areas remains in question, since with sufficiently large populations, even tiny differences will produce a statistically significant result. Shwartz et al. [13] also noted the strong influence of the statistical methods used on the variation found in hospitalization rates. However, regardless of the method used, the same geographic areas were always found to have higher than expected hospitalization rates over the 3-year period of the study.

Another way to prove the reliability of data on variation in the use of health care facilities is to measure whether the differences remain stable over time. Wolfe et al. [14] investigated hospital discharge rates in 60 small communities in Michigan over a 5-year period. The ranking of these communities by discharge rates (both surgical and non-surgical) remained rather equal. Therefore, there was no regression to the mean and these data confirm the existence of small area variation.

Taken together, differences in the use of health care services exist on all geographic levels. However, in analysing area variation, not all differences will prove to be statistically significant and much variation may be largely due to chance.

2.3. Potential determinants

Theoretically, four groups of determinants of differences in the use of health care services can be discerned (Table 1). The first group is related to the characteristics of the patient. Not only does this include the case-mix but also the degree of consumerism of the population. In small area analysis, differences in production rates between providers will, as far as is possible, be adjusted for these patient characteristics. However, not all the characteristics which can influence the use of health care services will be known. The second group of determinants are linked to characteristics of the providers, in this case, the physicians. Examples of these include the method of reimbursement (whether fee-for-service or salaried), their specialization, their experience and the number of years in practice, all of which can have an influence on the practice style of the physician. These determinants can result in divergent indications of diagnostic and therapeutic procedures with a possible consequence being a difference in the quality of care. A third group of determinants emanates from the institution in which the provider practices and includes such factors as the amount of facilities available (beds, diagnostics) and the way the care has been organized (group or solo practice, partnerships). Lastly, the environment of the providers can also influence the use of health care facilities. Examples here include the hospital referral practice of general practitioners or the physical distance to the institution (degree of urbanization).

This composition will be illustrated with results from recent literature, some of which have already been mentioned in the previous section on the recent evidence

Table 1
Determinants of differences in the use of health care facilities

	Determinant
Patient	Case-mix: Morbidity Socio-economic status Level of education Consumerism
Provider	Practice style: Method of reimbursement Experience
Institution	Facilities Practice organization Teaching status
Environment	Referral practice of general practitioners Urbanization

of variation. Each of the determinants can of course be responsible for the differences between areas and between providers, but the relative contribution of these factors to these differences in relation to the geographic level studied is more important. In the following survey, the determinants are discussed separately but, when known, the interdependent influence of these determinants will be mentioned.

2.3.1. Patient characteristics. In a study into the variation in age-adjusted discharge rates for musculoskeletal diseases between hospital market communities in Michigan, 26.6% of this variation was explained by socio-economic variables which had remained unchanged over an 8-year period [15]. These communities range in population from 9108 to 839 410 with a mean of about 150 000. In an elegant study of patient and physician contributions to the variations in cataract surgery rates across U.S. metropolitan areas, the influence of patient care-seeking behaviour appeared to be the largest contributory factor, and not the practice style of the physician [16]. In the research on antibiotic use in patients with pharyngitis between two small areas mentioned earlier, using logistic regression analysis, patients' clinical characteristics were predictors of treatment, and not the site of the HMO [5]. In a recent study into variations in acute cardiac ischaemia admissions in northern Michigan, two demographically practically identical towns appeared to differ by a factor of three; in the high-admitting hospital nearly twice as many patients were presented than in the low-admitting hospital [17].

2.3.2. Physician characteristics. In the Netherlands, Bensing et al. [18] found that female general practitioners and practitioners who worked part-time prescribed less drugs, but ordered more laboratory tests, than did their male and full-time colleagues. In this study, substitution between different practice activities appeared to exist. In another study, female physicians spent more time on preventing services [19]. The influence of gender on the provision of health care was also found in an Italian study: the more female surgeons surgical centres had, the more breast-conserving surgery was performed [20]. In this case, gender differences between groups of physicians can result in practice variation on a small geographic level.

In three major U.S. cities, physicians who practised solo and were paid fee-for-service had 41% more hospitalizations than did salaried physicians in health maintenance organizations [21]. In the above-mentioned study on geographic variation in expenditures for physicians' services in 317 U.S. metropolitan areas, the expenditures were not related to the number of physicians per capita but to the percentage of primary care physicians: the higher this percentage, the lower the expenditure [2]. In a study which tried to explain variations in length of stay in an university teaching hospital, attending and resident physicians accounted for only a small amount of these variations [22]. In research into geographic variation in the utilization of cataract surgery, in which the U.S. was divided in 181 regions, no association was found with the number of ophthalmologists but with the concentration of optometrists instead [23].

2.3.3. Hospital characteristics

In their study into differences in hospital readmission rates between Boston and New Haven, the researchers found that these differences could not be explained by the severity of the illness and suggested that the wider availability of hospital beds in Boston was the reason for these differences [4]. In Ontario, Canada, variations in the length of stay following acute myocardial infarction were explained by the presence of particular patient and hospital characteristics for only 12%, although a correlation was found between a lower hospital case-load and increased length of hospitalization [24].

2.3.4. A case of benign prostate hyperplasia

In our own study on variations in treatment of patients with benign prostate hyperplasia, the preliminary results show that various determinants of the differences could be identified. For a patient aged between 65 and 74 without comorbidity and with only slight irritative and obstructive complaints (a 'mean' patient), the chances of undergoing ultrasonography of the prostate were 58%. With more irritative complaints however, this decreases to 45%. If a 'mean' patient is treated by a urologist with less than 5 years' experience, the chances of this diagnostic procedure being implemented rise to 53%. If the specialist has been practising for more than 10 years this chance decreases to 39%.

Comparable variation appeared to exist in the choice of the initial treatment. With low irritative complaints, the chance of watchful waiting instead of surgery or drug treatment is 63.8%, while with high irritative complaints the patient only has a 29.6% chance of receiving no active treatment. A urologist with less than 5 years' experience will advise watchful waiting for 29.7% of his patient population, a specialist with 10 or more years experience applies this type of treatment for 35.4% of his population.

In conclusion, various types of independent variables determine variations in the use of health care facilities. However, more research is needed on the relative contribution of each of the determinants; for example, the degree of consumerism in relation to the various types of procedures, as has been investigated in cataract surgery. This research needs to be conducted on different levels; from the individual physician to large geographic areas, because on each level other determinants will play a role. On the individual level, this will be the practice style of the physician which is determined by his education, his experience and the method of reimbursement. On the institutional level, the setting of care provision will be probably more important while, on a larger scale, the morbidity of the population or the health care system is the putative determinant. Although in most studies some adjustment for case-mix is applied, minor differences between populations are mostly not taken into account. The question that has to be addressed is whether differences in quality of care is one of these determinants. Although Greenfield et al. [21] found a greater incidence of hospitalizations in both solo and fee-for-service practices than in HMOs, this does not necessarily imply that there is a difference in the quality of care. To answer this question, more research has to be conducted into the appropriateness of use of diagnostic and therapeutic procedures.

3. Assessment of appropriate use

3.1. The methods employed

Over the past 15 years, many studies have been conducted in which the appropriate use of diagnostic and therapeutic procedures have been assessed. Among the methods employed in this research, the modified Delphi method developed at the Rand Corporation by Brook and coworkers is the most well-known [25]. It will be described in detail because with this method the most experience has been gained in studying the relationship between practice variation and appropriateness of care. There are of course other ways to assess the appropriateness, for example by guidelines developed by the Agency for Health Care Policy and Research, and by the evidence-based medicine approach of the Cochrane Collaboration. In the Rand studies, the use of a procedure is defined as appropriate when the benefits (in terms of mortality and morbidity) exceed the risks when compared to other procedures, or with doing nothing. Monetary cost considerations are excluded. Before going into this method, and the results of the Rand studies, some general remarks on assessment procedures will be made. As will be shown, the type of method used can influence the result of the assessment.

Three steps can be discerned in the assessment procedure: the criteria to be assessed, the data that will be used, and the comparison of criteria and data (Table 2). Criteria can be either implicit or explicit. Explicit criteria can be developed by an external party (financiers, patient organizations), by an expert panel or by the caregivers themselves. The data used to judge the appropriateness can be the medical record or abstracts derived from this record, as is done in retrospective studies, or it can be gathered on a separate form in prospective designs. The judgement itself can be done by the external party or the expert panel who have established their own criteria, or by independent people, such as the researchers who make use of explicit criteria. Lastly, the caregivers themselves can judge their own performance as happens in some forms of peer review.

Table 2
Components of the assessment procedure

	Component
Criteria	Implicit versus explicit
	Established by experts
	external party
	caregivers themselves
Data	Medical record
	Abstract from medical record
	Separate form
Judgement	Experts
	External party
	Caregivers themselves
	Researchers

3.2. The Rand method

Within this method a very structured approach is followed. Firstly, a literature review is done on the efficacy and safety of the procedure. All the known indications in the form of typical patient groups are described. In the second phase, a multispecialty group of physicians is given the review and the indication list and are asked to rate the appropriateness of the procedure for each indication (1 = low to 9 = high). During a 2-day meeting, the ratings of each expert are tabulated and feedbacked with the group's distribution at the ratings. The panel discusses the ratings and all members rate anew the appropriateness. All indications with a mean rating of between 7–9 are appropriate, 4–6 are uncertain and 1–3 are inappropriate. Finally, these appropriate ratings are applied by independent researchers to a set of abstract data from the medical records. The abstract from the medical record has been prepared by trained nurses.

Over the past 7 years, the appropriateness of a number of procedures has been assessed and the results published. Table 3 gives an overview of these results [25–30]. As the table shows, the procedures assessed have been performed on an inappropriate indication in rather high percentages: it considers both diagnostic and therapeutic procedures. Interestingly, the appropriateness rate for coronary artery bypass graft surgery increased between 1988 and 1993 from 65 to 91%. However, these two studies were conducted in different areas of the U.S. [26,30].

Some international comparative studies using the Rand method have also been performed. In such studies, not only can the appropriateness rates of procedures performed in different countries be compared, but the rating of panels from these countries can also be set side by side. Table 4 shows three of these studies. In the study of Brook et al. [31], two panels, one from the U.S. and one from the UK, developed their judgement criteria by which the appropriate use of the procedure performed in various patient groups from the U.S. was assessed. The U.S. panel judged more indications appropriate than did the UK panel. In the study of McGlynn et al. [32], two different panels and two different patient groups from the U.S. and Canada were compared. In their study the criteria of the U.S. panel were also more liberal; the appropriateness ratings of U.S. and Canadian patients were comparable, however, Bengtson et al. [33] in Sweden used a modified Rand method

Table 3
Recent results of appropriateness assessment using the Rand method

Authors	Procedure	Appropriate (%)	Uncertain (%)
Chassin et al. [25]	Coronary angiography	74	9
	Carotid endarterectomy	35	32
	Gastrointestinal tract endoscopy	72	14
Winslow et al. [26]	Coronary artery bypass	65	30
Bernstein et al. [27]	Hysterectomy	58	25
Bernstein et al. [28]	Coronary angiography	76	20
Hilborne et al. [29]	PTCA	58	38
Leape et al. [30]	Coronary artery bypass	91	7

Table 4
Recent international comparative studies using the Rand method

	Appropriate use (%)	
Brook et al. [31]	U.S. panel	UK panel
Coronary angiography	87	65
Coronary artery bypass	73–83	40–58
McGlynn et al. [32]	U.S. panel	Canadian panel
Coronary angiography		
Canadian patients	76.7	57.8
U.S. patients	75.7	50.7
Coronary artery bypass		
Canadian patients	88.4	85.0
U.S. patients	90.6	84.5
Bengtson et al. [33]; Bernstein [34]	U.S. criteria	Swedish criteria
Coronary angiography		
Swedish patients		89
U.S. patients	73	49
Coronary artery bypass + PTCA		
Swedish patients		91
U.S. patients	93	57

and found a rather high rate of appropriate use of the two procedures in their country. In a letter to the Editor, Bernstein from the Rand group [34] stated that when he applied the Swedish appropriateness criteria to New York State patients, he found significantly less appropriate care than was reported for Swedish patients. Using his own U.S. criteria, the appropriate ratings were much higher.

3.3. Methodological pitfalls of assessment procedures

As mentioned earlier, the assessment of appropriateness can be performed by using either implicit or explicit criteria. Both types of criteria have their advantages and drawbacks (Table 5). With implicit criteria, the judgement will probably be more valid as the relevant care that has to be assessed will be judged. However, reliability can be low because how the judgement was arrived at is not verifiable. With explicit criteria, only that part of care can be judged for which criteria have

Table 5
Advantages and drawbacks of criteria used in the assessment procedure

Implicit criteria	Explicit criteria
Subjective	Objective
Valid	Reliable
Total care	Part of the care
Flexible	Rigid
Experts	Lay-persons
Time-consuming	Quick

been developed. With implicit criteria, the total care can be considered and individual circumstances can be taken into account. Lastly, explicit criteria can be applied by lay-persons and applying them will be less time-consuming than using implicit criteria. These criteria can only be used by experts.

In a review of the literature on results of implicit evaluation of patient care between 1966 and 1990 for studying interviewer agreement, Goldman [35] could identify 12 studies with enough data to analyze. Only two of these 12 studies had fair to good agreement between the reviewers with kappa of 0.4 and higher. In line with these findings, Hayward et al. [36] found a low degree of agreement in an implicit review of 675 patients records by 12 trained internists. Only the assessment of the overall quality of care, and of preventable death, reached an agreement of 0.5. For evaluating the appropriateness of hospital resource use, this type of review was unreliable.

Furthermore, judgement of the process of care by implicit criteria can be influenced by knowing the outcome. Caplan et al. [37] asked 112 anesthesiologists to judge the appropriateness of care in 12 cases. Every case was randomly presented with either a temporary or a permanent adverse outcome. The cases in which the outcome was stated as permanent were judged lower for appropriateness of care.

But also judgements formed using explicit criteria, as used in the Rand method, are ultimately determined by the experts who have developed them. Therefore, the composition of the panel is of utmost importance. A panel composed of performers of the procedure considered more indications appropriate for that procedure than did the physicians who referred the patients [38,39].

Fundamental criticism on methods in which consensus-based definitions of appropriateness are used, such as in the Rand method, has been expressed by Phelps [40] and Hicks [41]. Besides the bias of the literature review and the effect of the panel composition mentioned above, an appropriateness assessment has the same limitations as does any diagnostic procedure. For example, with high rates of appropriateness, a method that yields false positive measures of inappropriate use, will have a low predictive value. If this is the case, an appropriate intervention is labelled as inappropriate, something which is particularly troublesome on the level of the individual physician.

In most procedures of quality assessment, the process of care has been judged. This is the case in the appropriateness ratings of the Rand method. Over the last few years, health care has seen more emphasis being placed on outcome assessment. The question which has then to be addressed is whether there is any correlation between the assessment of process and outcome of care. In their classic study from 1973, Brook and Apple [42] went into this question. In this study, the care of 296 patients with urinary tract infection, hypertension or duodenal lesions was reviewed using implicit and explicit criteria and both process and outcome were assessed. Judgement of process using explicit criteria yielded the fewest acceptable cases (1.4%) while implicit judgement of outcome in 63.2% of cases was rated as adequate. In a study of Nobrega et al. [43], the process and outcome of care of 138 patients with hypertension was evaluated separately using explicit criteria: no significant association was detected between the quality of process and outcome.

In conclusion, many remarks can be made on the various quality assessment procedures. The result of the assessment procedure is dependent upon the method used, for example, whether implicit or explicit criteria are used. Judgement based on implicit criteria only, seems to be very unreliable. However, explicit criteria as developed by the Rand method also has its shortcomings: these criteria may be biased, as is shown by the influence of the panel composition and the difference between countries. Therefore, further study is necessary into methods of assessment of appropriateness to develop more valid and reliable instruments. Examples of research topics include comparison between the use of implicit and explicit criteria in order to develop a method that combines the strengths of both approaches. Furthermore, one must question the value of process assessment where no clear relationship appears to exist between the degree to which process criteria have been met, and the outcome of care. In this regard, follow-up is needed to establish what has happened to patients who underwent a diagnostic or therapeutic procedure that was designated as inappropriate. In case of performing an inappropriate diagnostic procedure: was the therapeutic intervention described and, if so, what was the outcome? In a study in the UK on appropriateness of care of coronary angiography, 21% of the procedures was judged as inappropriate; nevertheless, 38% of this patient group was operated [44]. The same question concerning the outcome can be addressed after a patient underwent an inappropriate therapeutic intervention. Answering these questions will give some insight into the validation of the method of appropriateness assessment.

4. Variation in the use of health care services and appropriateness rates

There are just a few studies in which variation in production rates are related to appropriateness assessment. In this section these studies will be considered together with the results of our own research in this field, to highlight the problems of such studies. As yet, no evidence has been found to suggest a correlation between the rate of procedures and their appropriateness. Therefore, in circumstances of higher use the degree of appropriateness, and thus the quality of care, is not necessarily lower.

This also holds true for the referral practice of general practitioners in which variation between individual providers is at stake. A referral can be considered as a (either diagnostic or therapeutic) procedure. Using implicit criteria, Knottnerus et al. [45] judged the quality of care of 192 referrals of four general practitioners in the Netherlands, two with high and two with average rates of referral to departments of internal medicine. In both groups, adjusted for age and sex, the appropriateness of referrals was the same: 57% in the high and 55% in the low referral general practitioners. Fertig et al. [46] came to the same conclusion in the UK. They found a 2.5-fold variation in referral rates among general practitioners to a particular hospital. However, elimination of all possible inappropriate referrals (12–20%, based on locally determined guidelines) could reduce the variation only to 2.1-fold.

In a recent study from the U.S. in which 2024 outpatient medical records of 135 providers in Maryland from different types of primary care settings were reviewed

Table 6
Appropriateness of use and production rates of three procedures

	Appropriate (%)	Equivocal (%)	Number of procedures per 10 000
Coronary angiography	72	10	50
	81	4	22
Carotid endarterectomy	37	34	23
	42	29	6
Upper GI tract endoscopy	71	11	149
	72	14	100

Data derived from Chassin et al. [25].

against explicit criteria, no consistent relation between appropriateness ratings and the use of resources could be found [47].

The most well-known study into a possible relationship between rates of use and appropriateness on the level of rather larger areas has been conducted by researchers from the Rand Corporation and the University of California at Los Angeles [24,48]. From their data of small area analysis, five geographic areas of high, average and low use of coronary angiography, carotid endarterectomy, and upper gastrointestinal tract endoscopy were selected and compared to levels in appropriateness. Only small differences in these levels could be found and the high-use site has the same appropriate rate as the low-use site [24] (Table 6). In one state, the same data were collected for 23 adjacent counties because it was assumed that, on this level, morbidity will not differ very much and the influence of practice style of groups of physicians could be more prominent [48]. However, only for coronary angiography did inappropriate use account for a 28% variance in this county state. For the other two procedures, no correlation was found on this area level.

In a special issue of *Health Services Research*, the findings of these two studies were discussed at length. Davidson focussed his criticism upon the low amount of empirical data that did not possess high enough levels of statistical power to provide an adequate test of the hypothesis, particularly in the study of one area [49]. Some procedures were not performed at all in some counties (see also the comments of Cain and Diehr [12]). Furthermore, it must be realized that at the same relative levels of inappropriate use, the providers with high production rates do have more inappropriate use in absolute terms! In fact, appropriate ratings should be calculated on the number of inhabitants in the region instead of on the number of procedures performed. Because the procedure was the start of the assessment, only potential over-use could be established, so potential under-use could not be detected.

4.1. A study of gonarthroscopies

In all studies on appropriate use, the procedures themselves, and not the clinical problem, were the subject of investigation, meaning that potential under-use could

not be established. We undertook a study of whether or not a specific procedure was carried out on patients with a defined clinical problem. This study will be discussed here to illustrate how such an investigation can be conducted and what problems were encountered. We decided to study gonarthroscopies performed by orthopaedic surgeons because this procedure is applied on a rather broad scale, is generally only carried out by these specialists in the Netherlands, and is reimbursed individually. Furthermore, in a pilot study regional differences regarding this procedure had already been established.

From seven of the 27 health regions in the Netherlands, all orthopaedic surgeons from the 33 partnerships were asked to log all patients aged between 15 and 45 with knee problems lasting more than 6 weeks during two separate 2-week periods in 1991 and 1992. On a separate form they were requested to supply a number of patient and disease characteristics, and to state whether or not an arthroscopy was performed.

An expert committee of the National Society for Orthopaedic Surgeons formulated guidelines for the patient group defined. These guidelines were approved with some minor changes after extensive deliberation during the annual meeting of the Society. The appropriateness of the decision whether or not to perform an arthroscopy was assessed by the researchers using the consensus guidelines.

During the two 2-week periods, 1221 patients were logged by 28 of the 33 partnerships of orthopaedic surgeons in the seven regions. However, 589 patients had to be excluded due to their complaint, or their simply not meeting the inclusion criteria. From the remaining 632 patients, a further 16 had to be excluded due to lack of data concerning the medical decision as regards treatment. Therefore, the definite number of patients available for analysis was 616; 305 of these patients underwent the procedure, while in 311 cases it was deemed unnecessary.

Table 7 shows the appropriateness rates in the seven regions. The mean true positive rate (appropriate to perform the procedure) was 83.0% and the mean true negative rate (appropriate not to perform the procedure) was 80.7%. There is, therefore, clear evidence of both over-use and under-use of gonarthroscopy in

Table 7
Appropriateness of the medical decision to perform an arthroscopy on patients aged between 15 and 45 with chronic knee complaints

Region	True positive	True negative
1	69.7 (33)	82.2 (45)
2	67.6 (34)	85.7 (28)
3	87.5 (88)	83.0 (49)
4	94.1 (17)	90.9 (11)
5	73.5 (34)	69.2 (13)
6	90.0 (50)	73.8 (61)
7	89.8 (49)	81.4 (59)
Total	83.0 (305)	80.7 (311)

Values are percentages; values in parentheses are the actual number of patients studied.

Table 8
Production rates of arthroscopies performed by orthopaedic surgeons in five of the seven regions in patients aged between 15 and 45 in relation to the true positive and true negative rates of appropriateness

Region	Number of athroscopies per 1000 inhabitants per year (aged 15–45 years)	True positive	True negative
1	6.07	69.7	82.2
2	4.22	67.6	85.7
3	5.29	87.5	83.0
5	7.60	73.5	69.2
7	4.26	89.8	81.4

typical orthopaedic practices. Using multiple logistic regression, it was found that the region itself could not account for differences in appropriateness. During 1992, all arthroscopies performed by orthopaedic surgeons could be collected for five of the seven regions. Table 8 shows the procedure rates, and the appropriateness assessment. From this table there seems to be no relationship between the number of arthroscopies and the appropriateness rates. However, due to both the small number of patients and regions, no statistical analysis could be performed. As already mentioned, this is a frequently occurring problem in this type of study. Furthermore, because of the required anonymous registration, we could not check up on the consecutiveness nor completeness of the data. Although we could calculate that we had missed a substantial amount of suitable patients, there was no indication of conscious selection by the medical specialists. More probably, due to the daily pressure of work, they forgot to register some suitable candidates for inclusion.

To sum up, no evidence has been found that practice variation is explained by differences in appropriateness rates. However, the studies concerned show methodological limitations and shortcomings; this conclusion is, so far, not definitive and more research is needed.

5. Conclusions: the agenda for further research

From this survey of the available research, the conclusion can be drawn that, up until now, there is no evidence to suggest that differences in appropriateness ratings can explain variation in the use of health care services or practice variation. This conclusion has two messages. The political message is that as yet high use of facilities does not necessarily mean providing care of low quality. The scientific message is that more research is warranted because the conclusion that this relationship does not exist in reality is still not a conclusive one.

Therefore, research is needed in three main areas: analysis of practice variation; assessment appropriateness; and the relationship between practice variation and appropriateness of care.

Firstly, analysis of differences in the use of health care services or practice variation has to be done with a large enough number of procedures, and has to be conducted on the various levels of care provision: from the individual physician to large geographic areas. In such analysis, on all levels, all potential factors that can be responsible have to be taken into account so that the relative contribution of each determinant on the various levels can be determined. In these studies, not only the supply of care has to be investigated but also the demand part, such as the degree of consumerism of the population.

Second, a more valid and reliable instrument for assessing appropriateness has to be developed. One explanation for not finding inappropriate use as a determinant of practice variation is that small differences in inappropriateness do exist but are not large enough to be detected with the current insensitive assessment methods. Subtle differences in practice style between physicians, such as a tendency towards performing new technologies or to apply procedures in a routine-based way without conscious decision making (enthusiasm versus uncertainty hypothesis), can produce large differences in the number of particular procedures within the boundaries of acceptable quality of care. In view of the shortcomings of the assessment of appropriateness with both implicit and explicit criteria, we need an instrument that combines the strength of both approaches. As a first step, further study that compares the use of implicit and explicit criteria should be conducted.

Furthermore, follow-up is needed of patients who underwent a diagnostic or therapeutic procedure that was designated as inappropriate in order to validate the assessment method.

Thirdly, to study the relationship between practice variation and appropriateness, the starting point has to be the clinical problem, otherwise potential under-use cannot be established. In addition, by departing from the clinical problem, substitution between various types of diagnostic or therapeutic procedures as reason for variation can be found. As stated above, high producents performed more inappropriate procedures in absolute terms than do their low producing colleagues. It is possible that substitution between different types of procedures occurs: a high rate of one particular procedure goes, in that case, together with a low rate of an alternative procedure for the clinical problem concerned. For example, general practitioners with a high referral rate to medical specialists probably prescribe fewer drugs, and both types of intervention may be appropriate for that clinical problem. However, in those studies in which some type of substitution could be measured, there was generally no evidence of this phenomenon.

Finally, we certainly will never find a complete explanation for the differences in use of health care facilities, but we must attempt to find as many causes as possible. However, because health care is provided by individual physicians to individual patients, some practice variation will always exist for the benefit of the patients.

References

[1] Lewis, C.E., Variations in the incidence of surgery, New England Journal of Medicine, 281 (1969) 880–885.

[2] Welch, W.P., Miller, M.E., Welch, H., Fischer, E.S. and Wennberg, J.E., Geographic variation in expenditures for physicians' services in the United States, New England Journal of Medicine, 328 (1993) 621–627.
[3] Welch, H.G., Miller, M.E. and Welch, W.P., Physician profiling. An analysis of inpatients practice patterns in Florida and Oregon, New England Journal of Medicine, 330 (1994) 607–612.
[4] Fischer, E.S., Wennberg, J.E., Stukel, T.A. and Sharp, S.M., Hospital readmission rates for cohorts of medical beneficiaries in Boston and New Haven, New England Journal of Medicine, 331 (1994) 989–995.
[5] Poses, R.M., Wigton, R.S., Cebul, R.D., Centor, R.M., Collins, M. and Fleischli, G.J., Practice variation in the management of pharyngitis: the importance of variability in patients' clinical characteristics and in physicians' responses to them, Medical Decision Making, 13 (1993) 293–301.
[6] Report: Verschillen tussen gezondheidsregio's in gebruik van ziekenhuisvoorzieningen (Differences Between Health Regions in the Use of Hospital Facilities), KNMG, Utrecht, 1989.
[7] Casparie, A.F., Post, D., van Harten, W.H. and Gubbels, J.W., Differences in production between medical specialists. An inventory based on claim data to identify potential areas for quality-improvement activities, European Journal of Public Health, 3 (1993) 292–295.
[8] Casparie, A.F., Geschiere, R., Post, D. and van Harten, W.H., Verschillen in zorgverlening tussen 11 KNO-maatschappen in 2 ziekenfondsregio's (Differences in provision of health care between 11 ENT-partnerships in two intake areas of obligatory medical insurance organizations), Nederlands Tijdschrift voor Geneeskunde, 135 (1990) 754–758.
[9] Diehr, P., Small area statistics: large statistical problems, American Journal of Public Health, 74 (1984) 313–314.
[10] Diehr, P., Cain, K., Connell, F. and Volinn, E., What is too much variation? The null hypothesis in small-area analysis, Health Services Research, 24 (1990) 742–771.
[11] Diehr, P., Cain, K.C., Kreuter, W. and Rosenkranz, S., Can small-area analysis detect variation in surgery rates? The power of small-area variation analysis, Medical Care, 30 (1992) 484–502.
[12] Cain, K.C. and Diehr, P., Testing the null hypothesis in small area analysis, Health Services Research, 27 (1993) 267–294.
[13] Shwartz, M., Ash, A.S., Anderson, J., Iezzoni, L.I., Payne, S.M.C. and Restuccia, J.D., Small area variations in hospitalization rates: how much you see depends on how you look, Medical Care, 32 (1994) 189–201.
[14] Wolfe, R.A., Griffith, J.R., McMahon, L.F., Tedechi, P.J., Petroni, G.R. and McLaughlin, C.G., Patterns of surgical and nonsurgical hospital use in Michigan communities from 1980 through 1984, Health Services Research, 24 (1989) 67–82.
[15] McMahon, L.F., McLaughlin, C.G., Petroni, G.R. and Tedeschi, P.J., Small area analysis of hospital discharges for musculoskeletal diseases in Michigan: the influence of socioeconomic factors, The American Journal of Medicine, 91 (1991) 173–178.
[16] Escarce, J.J., Would eliminating differences in physician practice style reduce geographic variations in cataract surgery rates? Medical Care, 31 (1993) 1106–1118.
[17] Green, L.A. and Becker, M.P., Physician decision making and variation in hospital admission rates for suspected acute myocardial ischemia. A tale of two towns, Medical Care, 32 (1994) 1086–1097.
[18] Bensing, J.M., van den Brink-Muinen, A. and de Bakker, D.H., Gender differences in practice style: a Dutch study of general practitioners, Medical Care, 31 (1993) 219–220.
[19] Bertakis, K.D., Helms, L.D., Callahan, E.J., Azati, R. and Robbins, J.A., The influence of gender on physician practice style, Medical Care, 33 (1995) 407–416.
[20] Grilli, R., Scorpiglioni, N., Nicolucci, A., Mainini, F., Penna, A., Mari, E., Belfiglio, M. and Liberati, A., Variation in use of breast surgery and characteristics of hospitals' surgical staff, International Journal for Quality in Health Care, 6 (1994) 233–238.
[21] Greenfield, S., Nelson, E.C., Zubkoff, M., Manning, W., Rogers, W., Kravitz, R.L., Keller, A., Tarlov, A.R. and Ware, J.E., Variations in resource utilization among medical specialties and systems of care. Results from the Medical Outcome Study, Journal of the American Medical Association, 267 (1992) 1624–1630.
[22] Hayward, R.A., Manning, W.G., McMahon, L.F. and Bernard, A.M., Do attending or resident physician practice style account for variations in hospital resource use? Medical Care, 32 (1994) 788–794.

[23] Javitt, J.C., Kendix, M., Tielsch, J.M., Steinwachs, D.M., Schein, O.D., Kobb, M.M. and Steinberg, E.P., Geographic variation in utilization of cataract surgery, Medical Care, 33 (1995) 90–105.
[24] Chen, E. and Naylor, C.D., Variation in hospital length of stay for acute myocardial infarction in Ontario, Canada, Medical Care, 32 (1994) 420–435.
[25] Chassin, M.R., Kosecoff, J., Park, R.E., Winslow, C.M., Kahn, K.L., Merrick, N.J., Keesey, J., Fink, A., Solomon, D.H. and Brook, R.H., Does inappropriate use explain geographic variations in the use of health care services. A study of three procedures, Journal of the American Association, 258 (1987) 2533–2537.
[26] Winslow, C.M., Kosecoff, J.B., Chassin, M., Kanouse, D.E. and Brook R.H., The appropriateness of performing coronary artery bypass surgery, Journal of the American Medical Association, 260 (1988) 505–509.
[27] Bernstein, S.J., McGlynn, E.A., Sin, A.L., Roth, C.P., Sherwood, M.J., Keesey, J.W., Kosecoff, J., Hicks, N.R. and Brook R.H., The appropriateness of hysterectomy. A comparison of care in seven health plans, Journal of the American Medical Association, 269 (1993) 2398–2402.
[28] Bernstein, S.J., Hilborne, L.H., Leape, L.L., Fiske, M.E., Park, R.E., Kamberg, C.J. and Brook, R.H., The appropriateness of use of coronary angiography in New York State, Journal of the American Medical Association, 269 (1993) 766–769.
[29] Hilborne, L.H., Leape, L.L., Bernstein, S.J., Park, R.E., Fiske, M.E., Kamberg, C.J., Roth, C.P. and Brook, R.H., The appropriateness of use of percutaneous transluminal coronary angioplasty in New York State, Journal of the American Medical Association, 269 (1993) 761–765.
[30] Leape, L.L., Hilborne, L.H., Park, R.E., Bernstein, S.J., Kamberg, C.J., Sherwood, M. and Brook, R.H., The appropriateness of use of coronary artery bypass graft surgery in New York State, Journal of the American Medical Association, 269 (1993) 753–760.
[31] Brook, R.H., Kosecoff, J.B., Park, R.E., Chassin, M.R., Winslow, C.M. and Hampton, J.R., Diagnosis and treatment of coronary disease: comparison of doctors' attitudes in the USA and the UK, Lancet, 1 (1988) 750–753.
[32] McGlynn, E.A., Naylor, C.D., Anderson, G.M., Leape, L.L., Park, R.E., Hilborne, L.H., Bernstein, S.J., Goldman, B.S., Armstrong, P.W., Keesey, J.W., McDonald, L., Pinfold, S.P., Damberg, C., Sherwood, M.J. and Brook, R.H., Comparison of the appropriateness of coronary angiography and coronary artery bypass graft surgery between Canada and New York State, Journal of the American Medical Association, 272 (1994) 934–940.
[33] Bengtson, A., Herlitz, J., Karlsson, T., Brandrup, G. and Hjalmarson, A., The appropriateness of performing coronary angiography and coronary artery revascularization in a Swedish population, Journal of the American Medical Association, 271 (1994) 1260–1265.
[34] Bernstein, S.J., Letter to the Editor: The appropriateness of coronary procedures in Sweden, Journal of the American Medical Association, 272 (1994) 1254–1255.
[35] Goldman, R.L., The reliability of peer assessments of quality of care, Journal of the American Medical Association, 267 (1992) 958–960.
[36] Hayward, R.A., McMahon, L.F. and Bernard, A.M., Evaluating the care of general medicine inpatients: how good is implicit review? Annals of Internal Medicine, 118 (1993) 550–556.
[37] Caplan, R.A., Posner, K.L. and Cheney, F.W., Effect of outcome on physician judgments of appropriateness of care, Journal of the American Medical Association, 265 (1991) 1957–1960.
[38] Scott, E.A. and Black, N., Appropriateness of cholecystectomy in the United Kingdom — a consensus panel approach, Gut, 32 (1991) 1066–1070.
[39] Fraser, G.M., Pilpel, D., Kosecoff, J. and Brook, R.H., Effect of panel composition on appropriateness ratings, International Journal for Quality in Health Care, 6 (1994) 251–255.
[40] Phelps, C.E., The methodologic foundations of studies of the appropriateness of medical care, New England Journal of Medicine, 329 (1993) 1241–1245.
[41] Hicks, N.R., Some observations on attempts to measure appropriateness of care, British Medical Journal, 309 (1994) 730–733.
[42] Brook, R.H. and Apple, F.H., Quality-of-care assessment: choosing a method for peer review, New England Journal of Medicine, 288 (1973) 1323–1329.

[43] Nobrega, F.T., Morrow, G.W., Smoldt, R.K. and Offord, K.P., Quality assessment in hypertension: analysis of process and outcome methods, New England Journal of Medicine, 296 (1977) 145–148.
[44] Gray, D., Hampton, J.R., Bernstein, S.J., Kosecoff, J. and Brook, R.H., Audit of coronary angiography and bypass surgery, Lancet, 335 (1990) 1317–1320.
[45] Knottnerus, J.A., Joosten, J. and Daams, J., Comparing the quality of referrals of general practitioners with high and average referral rates: an independent panel review, British Journal of General Practice, 40 (1990) 178–181.
[46] Fertig, A., Roland, M., King, H. and Moore, T., Understanding variations in rates of referral among general practitioners: are inappropriate referrals important and would guidelines help to reduce rates? British Medical Journal, 307 (1993) 1467–1470.
[47] Starfield, B., Powe, N.R., Weiner, J.R., Stuart, M., Steinwachs, D., Scholle, S.H. and Gerstenberger, A., Costs vs. quality in different types of primary care settings, Journal of the American Medical Association, 272 (1994) 1903–1908.
[48] Leape, L.L., Park, R.E., Solomon, D.H., Chassin, M.R., Kosecoff, J. and Brook, R.H., Does inappropriate use explain small area variation in the use of health care services?, Journal of the American Medical Association, 263 (1990) 669–672.
[49] Davidson, G., Does inappropriate use explain small-area variations in the use of health care services? A critique, Health Services Research, 28 (1993) 389–400.