Appropriate indications for prostatectomy in the UK – results of a consensus panel

D J W Hunter, C M McKee, C F B Sanderson, N A Black

Abstract

Study objective – The use of formal consensus development to determine appropriate indications for prostatectomy and to identify factors underlying clinical decisions about appropriateness is described.

Design – A nominal group technique was used.

Settings – The study took place in an academic research institution.

Participants – The panel consisted of six urologists and three general practitioners.

Measurements and main results – The panel identified agreed indications for prostatectomy, expressed in terms of different combinations of type of retention, type and severity of symptoms, and level of comorbidity. Agreement was reached for 67% of the indications considered. For acute or chronic retention, surgery is indicated, regardless of symptom severity, if life expectancy is greater than one year. For acute or chronic retention, surgery is generally indicated if symptoms are severe, or if symptoms are moderate and life expectancy is greater than five years. For patients with neither acute nor chronic retention, surgery is indicated if symptoms are severe, or if these are moderate and life expectancy is greater than five years. For chronic or acute retention surgery is inappropriate if symptoms are mild and life expectancy is less than one year, or if there is no retention and only mild symptoms. An “appropriateness score” was developed. This confirmed that in general the ratings were internally consistent, that the panel attached little weight to mild symptoms, that a combination of irritative and obstructive symptoms was no more indicative of surgery than obstructive symptoms alone, and that the type of symptom was less important than the other factors considered.

Conclusions – The results provide a basis for population based surveys of the need for prostatectomy.

(*J Epidemiol Community Health* 1994;48:58-64)

In the reformed National Health Service, health authorities are required to assess the needs of their populations and to purchase appropriate care. There is emerging agreement about the means to be used in assessing need or the requirement for specific health care inter-

ventions.¹⁻² The first step is to reach consensus, based on the best available evidence, about the appropriate indications for the intervention. The second step is to ascertain the population prevalence of individuals with appropriate indications. The final step is to determine which individuals with clinically appropriate indications actually want the intervention in question. Despite the importance of this type of research, relatively little has been undertaken.

Prostatectomy to remove obstruction of the urethra arising from benign prostatic hyperplasia (BPH) is one of the commonest elective surgical procedures. Patients commonly present with acute retention, chronic retention, or with a complex of obstructive and irritative symptoms such as hesitancy, poor flow, terminal dribbling, frequency, feeling of fullness, dysuria, and nocturia.³ Reported prostatectomy rates in western countries vary considerably, from 60 to 368 per 100 000.⁴⁻⁸ In general, the rates in North America are more than twice as high as in Europe. Even within the United Kingdom there is wide variation. The standardised prostatectomy ratio adjusted for age and private sector contribution varies in English regions from 77 to 144 (England = 100).⁹ There is little evidence that this is the result of differences in the prevalence of BPH, and much of the variation seems to result from the supply of surgical services, including the availability of surgeons and their judgement of the clinical appropriateness of surgery.^{8 10 11}

A review of published reports suggests only limited agreement about the appropriate indications for prostatectomy. For example, while some authors state that acute retention in the absence of detrusor failure or neuropathic obstruction is an absolute indication,¹² others argue in favour of conservative treatment for those who can void successfully after a period of catheter drainage.¹³⁻¹⁴ A recent study of the practice behaviour of urologists concluded that consensus about appropriate practice has yet to be achieved.¹⁵ A further controversy centres on whether it is better to perform surgery early, in order to avoid the greater risk when the patient grows older, or to limit surgery to those patients with symptoms that significantly affect their quality of life.¹⁶ These arguments have assumed greater importance after the publication of reports casting doubt on previous beliefs about the level of risk associated with transurethral resection of the prostate (TURP)¹⁷ and claims that up to 75% of TURPs performed in the United States may be inappropriate.¹⁸

Formal consensus development is a method of providing, on the best available evidence, an

Health Services
Research Unit,
Department of Public
Health and Policy,
London School of
Hygiene and Tropical
Medicine, Keppel
Street, London
WC1E 7HT
D J W Hunter
C M McKee
C F B Sanderson
N A Black

Correspondence to:
D J W Hunter.

Accepted for publication
June 1993

explicit basis for clinical decision making. One approach, the nominal group method, has been used increasingly in recent years to identify and resolve differences in clinical judgement. Developed by the Rand Corporation in the USA, it involves preparation of a review of published studies and two rounds (postal and face to face), in which a panel of relevant experts scores the indications for a particular intervention. The level of agreement is assessed and areas of consensus identified.¹⁹ Procedures that have been examined in this way in the UK include: coronary angiography, coronary artery bypass graft surgery, and cholecystectomy.^{20,21} This paper reports the use of a consensus panel to examine the appropriateness of the indications for TURP as the first stage in a study to assess the health care need for TURP for BPH in the population. A nominal group technique was used to determine the extent of agreement that existed on the appropriate indications. An appropriateness score was devised to examine the internal consistency of the group's opinions and the importance of various patient characteristics in decisions about the appropriateness of surgery.

Methods

A review of the published reports on the epidemiology of BPH, methods of treatment, indications, and outcome of prostatectomy was performed. Initially, MEDLINE was searched for the following medical subject headings (MeSH): prostatectomy, symptoms, standards, statistics and numerical data, and utilisation. The search covered the years 1983–91 and included papers in all languages.²² The review was supplemented by examination of major textbooks, follow up of additional references quoted in the papers located, and personal communication with other investigators known to be conducting research on this topic, both in the UK and the USA. In total, 175 papers were reviewed, of which 82 were included in the review of published reports.

The panel comprised five consultant urologists, one senior registrar in urology, and three general practitioners. In forming the panel, 13 consultant urologists in the North West Thames Regional Health Authority and 15 general practitioners, selected on the basis of previously expressed interest in the topic, were approached and asked to participate. The panel was drawn from teaching and non-teaching hospitals, as well as urban and rural districts. In the first round, panel members were sent the review of published reports and a questionnaire (developed from the review and with advice from a consultant urologist) that asked them to rate the appropriateness of potential indications for TURP. The indications were grouped into four categories of retention (acute, chronic, acute on chronic, and no retention), with differing types and levels of symptoms and different levels of comorbidity.

Chronic retention was defined as a post-void residual volume of 400 ml or greater. Sym-

toms were classified as irritative (frequency, nocturia, and dysuria); obstructive (hesitancy, poor flow, dribbling, and feeling of fullness); or a combination of both irritative and obstructive. Levels of symptom severity were categorised as none (symptoms never occur), mild (symptoms present occasionally or nocturia occurring once per night), moderate (symptoms present about half the time or nocturia two to three times per night), and severe (symptoms present most or all of the time or nocturia four or more times per night). Life expectancy was used as an indication of comorbidity. High comorbidity was defined as a life expectancy of less than a year, medium comorbidity as at least a year but less than five years, and low comorbidity as five years or more. It was assumed that there was no pre-operative evidence of prostatic cancer and that patients were fit enough to undergo general or spinal anaesthetic. Other forms of treatment such as drugs or stents were assumed to be available, so that if they were appropriate as first line therapy, TURP was deemed to be inappropriate.

Each combination of these four factors (type of retention, type of symptoms, severity of symptoms, and comorbidity) was taken to describe a category of patient. For example, one such category might represent patients in acute retention with mild obstructive symptoms and low comorbidity. If symptom severity was "none", further breakdown by symptom type was not applicable, and if there were no retention and no symptoms, further breakdown by comorbidity was not applicable. On this basis a total of 118 patient categories were assessed.

Appropriateness was scored using a 9 point scale, in which 1 indicated that the risk of TURP always outweighed the benefits and so surgery was always inappropriate, and 9 indicated that the benefits of TURP always outweighed the risks and so surgery was always appropriate. A score of 5 indicated that the risks and benefits were in balance.

After completing and returning the questionnaire, panellists attended a structured meeting facilitated by one of the authors (CMM), at which they were presented with a summary of the first round responses to the questionnaire. This indicated medians and ranges but not individual responses. The panellists discussed each patient category in turn and were given the opportunity to reconsider their previous responses. This final rating was used in the analysis.

Agreement was defined as when, after discarding the single highest and lowest ratings, the ratings of the seven remaining panel members fell within a 3 point range. Panellists were judged to have disagreed if, after discarding the highest and lowest ratings, at least one rating fell in the 1–3 range and at least one rating in the 7–9 range. Categories about which there was neither agreement nor disagreement were classified as subject to partial agreement. For patient categories about which the panel agreed, surgery was judged to be inappropriate if the median rating fell between

Table 1 Example of weights for dimensions of indications for surgery

Retention None: 0	Acute: 1	Chronic: 2	Acute on chronic: 3
Symptom severity None: 0	Mild: 1	Moderate: 2	Severe: 3
Symptom type Irritative: 0	Obstructive: 1	Irritative & obstructive: 2	
Comorbidity High: 0	Medium: 1	Low: 2	

Table 2 True and false positive and negative scores at lower and upper thresholds

Panel judgement	Appropriateness score			
	Below lower threshold	On above lower threshold	Below upper threshold	On above upper threshold
Surgery inappropriate	True - ve type 1	False + ve type 1	True - ve type 2	False + ve type 2
	False - ve type 1	True + ve type 1	True - ve type 2	False + ve type 2
Surgery indicated	False - ve type 1	True + ve type 1	False - ve type 2	True + ve type 2
	True - ve type 1	False + ve type 1	False - ve type 2	True + ve type 2

1–3 and appropriate if the median rating fell between 7–9. A rating of 4–6 was judged to be equivocal. The rationale for the use of this scoring system has been described elsewhere.²³ Data were analysed using programs written in the statistical program SAS.²⁴

The practical value of listing the appropriateness of each of 118 patient categories is limited and an attempt was made to summarise the results by devising a scoring system that could be used to estimate the appropriateness of surgery for any given patient category. With type of retention, symptom severity, symptom type, and comorbidity as the four factors that determine appropriateness, this involved assigning different weights to each “point” on four corresponding “dimensions”. An example of one such set of weights is shown in table 1. With this particular set, patients in acute retention with mild obstructive symptoms and low comorbidity would score 1 + 1 + 1 + 2 = 5.

The conclusions of the panel were divided into three groups: “appropriate”, “inappropriate”, and “unclear” where the panel either disagreed, agreed only partially, or were equivocal. The question was whether these conclusions were consistent with an underlying set of judgements about the relative importance in determining appropriateness of the four different “dimensions” and the various “scale

points” in each dimension. If so, it would be possible to devise a scoring system that could predict the panel’s ratings reasonably well. Patient categories with scores below a certain threshold would tend to have been rated by the panel as inappropriate for prostatectomy. Categories with scores above a second and higher threshold would tend to have been rated as appropriate. And categories with scores within a middle band between these two thresholds would tend to be those about which the panel results were unclear.

To examine this question the following procedure was used: (1) a set of weights was chosen; (2) a score was calculated for each category of patient; (3) a lower threshold value was chosen and rates of “type 1” true and false positives and negatives calculated, using the definitions shown in table 2; (4) an upper threshold value was chosen and rates of “type 2” true and false positives and negatives were calculated (table 2). A set of weights together with a lower and an upper threshold value thus constitute a scoring scheme.

This procedure was repeated for a variety of scoring schemes. In the main analysis only whole numbers were considered for the weights and threshold values. An intuitive set of weights was chosen as a starting point. First, a variety of ways of aggregating the weights in each dimension to form scores, including the sum, the product, and combinations (for example, symptom type × symptom severity + retention + comorbidity) was tried. Then a systematic procedure was followed to test the effect of giving more or less weight to each dimension and to selected scale points. Once a satisfactory weighting scheme had been found, the effect of various departures from it was examined. Over 40 schemes were tried in all. To assist in interpreting the results, an upper limit on the proportion of categories that could be correctly assigned by any scheme using the panel data (including one with non-integer weights and thresholds) was derived using conventional discriminant analysis. This produced two functions that divided the patient categories into inappropriate, unclear, and appropriate. Also it was necessary to take account of the fact that some categories will be correctly assigned by chance even if a scoring scheme has no real predictive value.

Figure 1 represents an “ideal” result with no false positives or false negatives at either threshold. In this hypothetical example, only the categories that the panel considered inappropriate scored less than 4, only the categories that they considered appropriate scored 6 or more, and only those about which the panel were unclear scored 4 or 5. This is only ideal given the observed distribution of panel ratings. Of course a result in which there were fewer categories for which the panel results were “unclear” would be better. Also it should be emphasised here that the figures given for the percentage agreement and disagreement, and for the percentage of false positive and false negatives, would only match those found in a survey of clinical practice if each category of patient were equally prevalent.

Figure 1. Prostatectomy indication score. Hypothetical ideal scheme.

Table 3 Final ratings of indications for transurethral resection of the prostate by the panel

	No	(%)
Agreement	79	(67)
Appropriate	53	(45)
Equivocal	6	(5)
Inappropriate	20	(17)
Partial agreement	32	(27)
Disagreement	7	(6)
Total	118	(100)

Results

PANEL RATINGS

The extent of agreement on indications is shown in table 3. The patient categories for which surgery was considered appropriate, unclear, or inappropriate are summarised in fig 2, in which no distinction is made between the type of symptoms. In general:

For acute on chronic retention, surgery is generally indicated regardless of symptom severity, if life expectancy is greater than one year;

For chronic or acute retention, surgery is indicated if symptoms are severe, or if symptoms are moderate and life expectancy is greater than five years;

Figure 2. Indications for prostatectomy.

For patients with neither acute nor chronic retention, surgery is indicated for patients with severe symptoms, or moderate symptoms and a life expectancy of greater than five years;

Surgery is inappropriate for chronic and acute retention if symptoms are non-existent or mild and life expectancy is less than one year, and for those with no retention and only mild symptoms.

There were a few apparent inconsistencies in the panel's judgements. For example, for a history of *either* acute *or* chronic retention with severe obstructive symptoms and high comorbidity, the panel were in agreement that prostatectomy was appropriate. However, for a similar case but with chronic *and* acute retention (a category which was, if anything, worse) the panel agreed only partially.

APPROPRIATENESS SCORES

Of the methods tried for combining weighted scores on each of the four "dimensions" to provide an overall score for each patient category, the simple sum performed as well as any and had the virtue of simplicity. The results in terms of the relative success of 10 illustrative sets of weights in predicting the panel's opinions are shown in table 4 and three are shown graphically in fig 3. In table 4 the three right hand columns give the percentage of all patient categories that were rated the same by the scoring system and the panel (true positives + true negatives), using the threshold levels shown in the two previous columns.

Given the numbers of categories identified by the panel as inappropriate, unclear, and appropriate, the proportion of "correct" assignments that could be expected from a scheme with no real predictive value (analogous to a system for predicting the results of tossing a coin that was only correct 50% of the time) would be 38%. The proportion of cases correctly identified in a two way split between (a) inappropriate and (b) unclear/appropriate would be 51%. The proportion correctly identified if the split were between (a) inappropriate/unclear and (b) appropriate would be 72%. The discriminant analysis "correctly" assigned 88% of the patient categories to inappropriate, unclear, or appropriate.

The starting point was scheme 1, an "intuitive" set of weights given in the example described in the methods section. Some members of the panel expressed the view that the difference between chronic and acute on chronic retention is unimportant in determining indications for surgery. Scheme 2 is consistent with this point of view, but it appears that in terms of its capacity to predict the panels' actual ratings it is slightly inferior to scheme 1. In scheme 3, no significance is attached to symptoms that are "only" mild. This is an improvement on scheme 1. In scheme 4, no distinction is made between irritative and obstructive symptoms. This is a slight improvement on scheme 3. In scheme 5, no distinction is made between acute and chronic retention, and greater weight is attached to comorbidity.

Table 4 Proportions of correct assignments to inappropriate, unclear, and appropriate using different sets of weights for type of retention, severity and type of symptoms, and degree of comorbidity

Scheme	Retention	Symptom severity	Symptom type	Comorbidity	T1	T2	Match using 3 groups: inappropriate v unclear v appropriate (%)	Match using 2 groups: inappropriate v unclear + appropriate (%)	Match using 2 groups: inappropriate + unclear v appropriate (%)
1	0123	0123	012	012	3	6	76.3	89.0	87.3
2	0122	0123	012	012	3	6	75.4	88.1	87.3
3	0123	0023	012	012	3	6	79.7	92.4	88.1
4	0123	0023	011	012	3	5	82.2	94.1	88.1
5	0223	0023	011	023	4	6	83.9	94.9	89.0
6	0224	0024	011	024	4	7	85.6	96.6	89.0
7	0224	0024	000	024	4	6	78.0	96.6	81.4
8	0222	0024	011	024	4	7	80.5	93.2	87.3
9	0222	0023	011	023	4	6	83.1	94.9	88.1
10	0000	0024	011	024	2	4	73.7	89.8	83.9

T1 = lower threshold, between inappropriate and unclear; T2= upper threshold, between unclear and appropriate.
Match = % of assignments that match the conclusions of the panel about whether appropriate, unclear or inappropriate.

In scheme 6, the weights given to retention, symptom severity, and comorbidity are all increased relative to symptom type. This is the best overall scheme using integer weights. Scheme 7 explores the extent to which the appropriateness of surgery can be determined without information about the type of symptoms. Plainly this is much worse than other schemes at identifying appropriate cases, but is as good as any at identifying inappropriate ones, and is still better overall than schemes 1 and 2. Scheme 8 explores the question of how far the appropriateness of surgery can be predicted with no information about type of retention – whether it is acute, chronic, or acute on chronic. Scheme 9 is a slightly better variant of this in which less weight is given to severe symptoms and low comorbidity.

Scheme 10 confirms that information about whether the patient is in retention or not is, as one might expect, an important factor in determining the appropriateness of surgery.

The analysis in table 4 makes no distinction between false positives and false negatives. This matter is addressed in table 5. If the scores were to be used as part of a screening procedure in primary care, it would be important to avoid false negatives. If the risks of surgery were generally substantial compared with the benefits, it would be important to avoid false positives.

One further consideration is that scheme 4 is one that avoids the situation in which patient categories which have been rated by the panel as appropriate have the same score as categories that have been rated as inappropriate. One can thus say with scheme 4 that groups scoring 3 or less include all those judged by the panel as inappropriate for prostatectomy, and that groups with scores of 4 or more include all those for which prostatectomy was judged appropriate.

Tables 4 and 5 suggest a number of points. Firstly, retention, symptom severity, and the degree of comorbidity were given roughly equal weight by the panel when rating patient categories, whereas the type of symptoms was much less important. Secondly, the presence of retention is important. The difference between acute and chronic retention is less so, and acute on chronic retention is a stronger indication than either acute or chronic retention alone. Thirdly, in effect the panel did not distinguish between patients with mild symptoms and patients with no symptoms at all. Fourthly, the panel considered obstructive symptoms to be more important in indicating surgery than irritative symptoms. And fifthly, the type of symptom was not important in determining the groups for which prostatectomy was inappropriate.

Discussion

Consensus development has been advocated as a means of making explicit the basis for clinical decisions. Although the nominal group technique of consensus development is well established, the development of a model to explain the ratings given by a panel is new. We believe that this approach has clarified much of the thinking behind the normally implicit decision

Figure 3. Prostatectomy appropriateness score.

Table 5 Rates of false positives and false negatives for different threshold values discriminating between inappropriate, unclear and appropriate, and different sets of weights for type of retention, severity and type of symptoms, and degree of comorbidity

Scheme	Type 1 (threshold for "unclear")						Type 2 (threshold for "appropriate")					
	1	2	3	4	5	6	3	4	5	6	7	8
1	% false +ve		60	30	10	0			37	12	3	
	% false -ve		1	8	23	45			0	13	43	
2	% false +ve		60	30	10	0		63	32	8	0	
	% false -ve		2	10	28	51		0	2	19	53	
3	% false +ve		35	10	0				26	8	3	
	% false -ve		2	12	30				2	19	49	
4	% false +ve		25	0				48	18	6	2	
	% false -ve		2	14*				0	4	30	60	
5	% false +ve		50	10	0			66	48	17†	6	
	% false -ve		0	4	15			0	2	4	32	
6	% false +ve		50	15	10	0			55	32	15	8
	% false -ve		0	1	11	26			0	2	6	34
7	% false +ve	80	15	15	0			72	32	32	8	
	% false -ve	0	1	1	26			0	2	2	34	34
8	% false +ve	80	50	15	10	0		66	48	25	11	5
	% false -ve	6	8	11	22	39		11	11	17	26	60
9	% false +ve	50	10	0				63	43	12	5	0
	% false -ve	2	6	18				0	2	11	40	60
10	% false +ve	70	40	30	15	0	58	37	22	8	2	
	% false -ve	2	4	13	27	41	0	4	13	42	49	

False - ves % type 1: denominator: number of patient groups judged by panel -> appropriate or unclear
numerator: subset of denominator indicated by score -> inappropriate
type 2: denominator: number of patient groups judged by panel -> appropriate
numerator: subset of denominator indicated by score -> unclear or inappropriate
False + ves % type 1: denominator: number of patient groups judged by panel -> inappropriate
numerator: subset of denominator indicated by score -> unclear or appropriate
type 2: denominator: number of patient groups judged by panel -> inappropriate or unclear
numerator: subset of denominator indicated by score -> appropriate
* Indicates a false negative rate of 14% for a type 1 threshold of 4. This implies that if a score of less than 4 is taken to mean that prostatectomy is inappropriate, then 14% of all indications judged by the panel as appropriate or unclear were indicated as inappropriate by the score.
† Indicates a false positive rate of 17% for a type 2 threshold of 6. This implies that if a score of 6 or more is taken to mean that prostatectomy is appropriate, then 17% of all indications judged by the panel to be inappropriate or unclear were indicated as appropriate by the score.
The results for thresholds at the upper and lower ends of the scale do not appear in the table because there are either no false positives or no false negatives.

to recommend prostatectomy, and we suggest that it could be usefully employed with the results of similar panels that have considered other procedures. The panel discussion was lively, with all members participating, and very little clear disagreement. In addition, as can be seen from fig 2, the ratings were generally internally consistent. This internal consistency may have been helped by the layout of the forms which the participants were asked to complete.

The panel did not consider that surgery was necessarily appropriate in patients with a history of acute retention, although there is a large group of these patients in whom surgery will sometimes be appropriate. There is also support for the approach described as "watchful waiting".

This study has highlighted a question that has received little attention in the published reports: do irritative symptoms have a different prognostic implication than obstructive symptoms? Some panellists felt that the exercise had involved an artificial distinction between obstructive and irritative symptoms which was seldom made in practice, whereas others expressed the view that prostatectomy should be performed for obstructive but not for irritative symptoms, because irritative symptoms were caused by detrusor instability rather than an enlarged prostate. The weighting analysis showed that the latter view was the more consistent with the panel's final judgements. Attaching equal weight to obstructive and to combined obstructive plus irritative symptoms but less weight to irritative symptoms alone gave the best "fit", but the type of symptom was the least important of the four "dimensions" considered by the panel. It also confirmed, despite the view of some that comorbidity is unimportant because of the ability

to perform prostatectomy under spinal anaesthesia, that comorbidity does play a part in determining appropriateness.

The weighting analysis also indicated that if the objective is simply to eliminate patients who are inappropriate for TURP, then although the presence of retention is important, the type of retention is less so. This suggests that it may be possible to devise useful guidelines for general practitioners about what kinds of patients to refer.

In considering these results, the limitations of consensus development techniques need to be remembered. Firstly, there is a potential bias in the selection of the panellists. We cannot be sure of the extent to which our panel reflected the views of other urologists and general practitioners. Although it has been shown that those doctors who are willing to participate in expert panels are representative of their colleagues,²⁵ the exact composition of a panel can effect the results obtained.²⁶ In this study, all the panel members came from one health region, in which the prostatectomy rate was 23% lower than the national average in 1985.⁷ It may well be that a panel consisting of consultants from other regions would have given different results. Secondly, as well as any selection bias, the results will also be affected by any "random" variation in panel behaviour. Ideally the whole process should be replicated using different panels.

Thirdly, the ratings of a panel are dependent on the definitions used. As in many areas of clinical work, terms in everyday use are used differently by different clinicians. Although there was some uncertainty among a few of the panellists at first, agreed definitions were developed and used as a basis for the second round, in some cases leading to modification of earlier scores. We do not know to what extent

panellists made use of the review of published reports that had been provided, although most stated that they had found it helpful.

Fourthly, we do not know how far stated opinions correspond to actual clinical practice, though this will be examined and reported in another paper.

Fifthly, the importance that should be attached to the figures given for the percentage agreement and disagreement, and to the percentage of false positive and false negatives, remains uncertain. As was pointed out earlier, they would only match those found in a survey of clinical practice if each category of patient was equally prevalent. In practice, some of the combinations of signs, symptoms, and comorbidity that the panel was asked to consider may be very rare. A proper interpretation of the importance of the extent of agreement over ratings would need to take the prevalence of each patient category into account, and the necessary data are not yet available. This is the subject of current research.

Finally, the existence of a consensus does not in itself mean the consensus view is correct and there is a danger that the nominal group process will arrive at collective ignorance rather than wisdom. As such it is not a replacement for either rigorous scientific reviews of published reports or for research, but rather a means of identifying current medical opinion and areas of disagreement. Ideally, indications for surgery should be based on studies of the relationship between patient characteristics and clinical outcome. To be reliable these need to be based on large numbers of patients and such data are beginning to become available.

How will the results of the study be used? In the immediate future, it will be used to assess the prevalence of individuals with appropriate indications for surgery among the population. The appropriateness score could be used to develop guidelines for general practitioners about the kind of patients to refer. Further work will examine its ability to predict outcome and the extent to which surgeons follow these policies in practice.

We are grateful to those who participated on the panel: Mr Allen, Mr B Ellis, Mr M Emberton, Dr T S V Bostock, Dr M Clayson, Dr N Rea, Mr R Shearer, Mr M Snell, Mr N Waterfall, Mr G Williams. We thank Mr Paul Abel for reviewing background material for the panel. We would like to thank Dr Sheila Adam of the Public Health Directorate, North West Thames Regional Health Authority for providing financial support for this project. Thanks are also due to Dr Aileen Clarke for reviewing a draft of this manuscript.

- 1 Frankel S. The epidemiology of indications. (Editorial). *J Epidemiol Community Health* 1991;45:257-9.
- 2 McKee M, Sanderson CFB. An approach to assessing need for elective surgery: the importance of thresholds. In: Ludicissa MS, Fayos RG, eds. *Proceedings of the European Healthcare Management Association Annual Conference, Madrid, 1991*. Madrid: Fundacion MAPFRE Medicina Escuela Nacional de Sanidad, 1991:359-67.
- 3 Doll HA, Black NA, McPherson K, Flood AB, Williams GB, Smith JC. Mortality, morbidity and complications following transurethral resection of the prostate for benign prostatic hypertrophy. *J Urol* 1992;147:1566-73.
- 4 Quill DS, Devlin HB, Plant JA, et al. Surgical operation rates: a twelve year experience in Stockton on Tees. *Ann Roy Coll Surg Eng* 1983;65:248-53.
- 5 Stephenson WP, Chute CG, Guess HA, Schwartz S, Lieber M. Incidence and outcome of surgery for benign prostatic hyperplasia among residents of Rochester, Minnesota: 1980-87. A population-based study. *Urology* 1991;38(1 suppl):32-42.
- 6 Department of Health, Office of Population Censuses and Surveys. *Hospital inpatient enquiry: inpatient and day case trends, 1979-85*. London: HMSO, 1989.
- 7 McPherson K. Why do variations occur? In: Mooney G, Anderson TF, eds. *The challenge of medical practice variation*. London: Macmillan, 1990.
- 8 Nicholl JP, Beeby NR, Williams BT. Role of the private sector in elective surgery in England and Wales, 1986. *BMJ* 1989;298:243-7.
- 9 Wennberg JE, Mulley AG, Hanley D, et al. An assessment of prostatectomy for benign urinary tract obstruction. *JAMA* 1988;259:3027-30.
- 10 Graversen PH, Gasser TC, Watson JH, Himman F Jr, Bruskewitz RC. Controversies about indication for transurethral resection of the prostate. *J Urol* 1989;141:475-81.
- 11 Blandy JP. The indications for prostatectomy. *Urol Int* 1978;33:159-70.
- 12 Birkhoff JD, Weiderhorn AR, Hamilton ML, et al. Natural history of benign prostatic hypertrophy and acute urinary retention. *Urology* 1976;7:48-52.
- 13 Spiro LH, Labay G, Orkin LA. Prostatic infarction: its role in acute urinary retention. *Urology* 1974;3:345-7.
- 14 Lloyd SN, Kirk D. Who should have a prostatectomy? A survey of the management of patients presenting with bladder outflow obstruction. *J R Soc Med* 1991;84:533-5.
- 15 Fisher ES, Wennberg JE. Administrative data in effectiveness studies: the prostatectomy assessment. In: Heithof KA, Lohr KN, eds. *Effectiveness and outcomes in health care*. Proceedings of an invitational conference by the Institute of Medicine. Washington: National Academy Press, 1990.
- 16 Roos NP, Wennberg JE, Malenka DJ, et al. Mortality and reoperation after open and transurethral resection of the prostate for benign prostatic hyperplasia. *N Engl J Med* 1989;320(17):1120-4.
- 17 Cotton P. Case for prostate therapy wanes despite more treatment options. *JAMA* 1991;266(4):459.
- 18 McGlynn EA, Kosecoff J, Brook RH. Format and conduct of consensus development conferences: multinational comparison. *Inter J Tech Assess Health Care* 1990;6:450-69.
- 19 Gray D, Hampton JR, Bernstein SJ, Kosecoff J, Brook R. Audit of coronary angiography and bypass surgery. *Lancet* 1990;335:1317-20.
- 20 Lomas J, Anderson MD, Enkin M, Vayda E, Roberts R, MacKinnon B. The role of evidence in the consensus process: results from a Canadian consensus exercise. *JAMA* 1988;259(20):3001-5.
- 21 US National Library of Congress. *MEDLINE on CD-ROM*, Silver Platter, Bethesda, 1992.
- 22 Scott EA, Black NA. When does consensus exist in expert panels? *J Public Health Med* 1991;13(1):35-9.
- 23 SAS Institute Inc. *Statistical Analysis System Version 5*. Cary, NC: SAS Institute Inc, 1989.
- 24 McKee M, Priest P, Ginzler M, Black N. How representative are members of expert panels? *Quality Assurance in Health Care* 1991;3(2):89-94.
- 25 Scott EA, Black NA. Appropriateness of cholecystectomy - a consensus panel approach. *Gut* 1991;9:1066-70.