Judgment Analysis of Surgeons' Prioritization of Patients for Elective General Surgery

Andrew D. MacCormick, MBChB, Bryan R. Parry, FRACS

***Background**. Access to elective general surgery in New Zealand is governed by clinicians' judgment of priority using a visual analog scale (VAS). This has been criticized as lacking reliability and transparency. Our objective was to describe this judgment in terms of previously elicited cues. **Methods**. We asked 60 general surgeons in New Zealand to assess patient vignettes using 8 VAS scales to determine priority. They then conducted judgment analysis to determine agreement between surgeons. Cluster analysis was performed to identify groups of surgeons who used different cues. Multiple regression for the combined surgeons was undertaken to determine the predictability of the 8-scale VAS. **Results**. Agreement between surgeons was poor (r_a = 0.48). The cause of poor agreement was mostly due to poor consensus (G) between surgeons in how they weighted criteria. Using cluster analysis, we classified the surgeons into 2 groups: 1 took more account of quality of life and diagnosis, whereas the other group placed more weight on the influence of treatment. The 8-scale VAS showed good predictability in assigning a priority score (R^2 = 0.66). **Discussion**. The level of agreement reflects surgeons' practice variation. This is exemplified by 2 distinct surgeon groups that differ in how criteria were weighted.*

Key words:** health care rationing; operative surgical procedures; waiting lists; judgment analysis; priority setting.* ***(Med Decis Making 2006;26:255–264)

In New Zealand, health services are predominantly state funded. Historically, due to many reasons including logistical factors, waiting lists for elective surgery have developed. In a move away from the waiting list system, the government instituted a "booking system" that required the provision of patient certainty of access within a 6-month period.[1] The booking system requires the triaging and subsequent denial of services to some patients who need surgery. This triage process is performed using algorithms or tools (clinical priority assessment criteria) that generate priority scores to indicate patients' priority for surgery.[2] Previously, the process had been performed by assigning arbitrary weightings to a priori determined criteria.

One such priority-scoring tool was the generic surgical priority criteria (GSPC). This tool was introduced into Auckland Hospital without prior consultation with clinicians. It was subsequently shown not to assign sufficient priority to patients with malignant disease.[3] A clinician-led initiative replaced the GSPC with a single visual analog scale (VAS), which was thought to reflect clinical judgment more appropriately. This single VAS provided priority scores that subsequently proved to be predictive of the time patients waited for surgery.[4] It, however, was not transparent as to how the priority score was determined (i.e., what criteria clinicians used to make to individual judgments).

This experience highlights different requirements held by 2 separate stakeholders. Clinicians want a priority-scoring tool to reflect clinical judgment, whereas policy makers want the prioritization process to be transparent so that patients and policy makers understand the processes used in clinical decision making. A synthesizing approach is to describe clinical judgment using social judgment theory (also known as judgment analysis or JAN).[5]

Received 23 May 2005 from the Department of Surgery, Faculty of Medical and Health Sciences, University of Auckland, New Zealand. Poster presented at the 24th Annual Meeting SMDM, Baltimore. Financial support for this study was provided in part by a contract with the Ministry of Health, New Zealand. The funding agreement ensured the authors' independence in designing the study, interpreting the data, and writing and publishing the report. Revision accepted for publication 10 November 2005.

Address correspondence to Dr. Andrew D. MacCormick, Department of Surgery, Faculty of Medical and Health Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand; e-mail: andrew.maccormick@auckland.ac.nz.

DOI: 10.1177/0272989X06288680

This body of theory requires 2 questions to be answered. The first is to determine the cues or criteria that surgeons believe they use in prioritizing patients for elective surgery. This has been performed previously and is reported elsewhere.[6] The second question is to determine how surgeons actually use these cues. This study addresses this latter issue.

Specifically our aims were to

1. identify individual surgeons' utilization of the cues in the prioritization of patients for elective general surgery,
2. identify the cause(s) of the hypothesized lack of agreement between surgeons and any remedial steps that may be taken to improve agreement,
3. identify any groupings of surgeons who use the cues in a different manner,
4. determine the predictability of a new 8-scale VAS priority scoring tool, incorporating the previous single VAS, to assign priority to patients.

This study received funding from the New Zealand Ministry of Health.

METHODS

Judgment Analysis

Theoretical Model

Social judgment theory is a conceptual form of judgment based on the analogy of a convex lens known as the Brunswik lens model (BLM).[5] In brief, within this model, the left "focal point" represents a "gold standard." The right focal point represents the subject's judgment. The set of cues is the "lens" that the subject—in our case, the surgeon—uses to make their judgment.

In this study, no gold standard was available because the true priority score was unknowable. In addition, we were more interested in the agreement between different surgeons than agreement with an arbitrary standard. A previous modification of the BLM allowed for the situation in which the comparison between subjects was of interest, and the gold standard was absent.[7] Thus, both the left and right focal point in our setting was a surgeon's judgment of a patient's priority for surgery.

Furthermore, we had 2 conceptual sets of cues. These were the "objective" cues present in the vignettes and the "subjective" interpretation of these cues by each surgeon as measured by 7 VAS (see below). The objective cues were those that were present in written form within the vignettes. They were not truly objective in that they were written descriptors of a patient. Thus, they did not include directly measurable factors. The subjective cues were the surgeons' interpretation of the value of the objective cues, within the vignette, measured on a 50-mm VAS. An expanded form of the BLM allows for at least 2 sets of cues.[8] This allowed us to represent the objective cues and their subjective interpretation. Therefore, a modified BLM was used (Figure 1). This was a synthesis of Brehmer's modification[7] and the expanded BLM developed by Stewart and Lusk.[8]

Seven previously elicited criteria were used as the cues for the surgeon's judgment of priority.[6] These included treatment, diagnosis, psychological impact of the diagnosis, patient characteristics, symptomatology, future complications, and impact on quality of life.

To achieve the first aim of our study, we needed to develop a model for each surgeon's judgment. This was performed with multiple linear regression, for individual surgeons, using the 7 cues as independent variables and the surgeon's judgment of patient priority, as measured by the 100-mm VAS (see below), as the dependent variable. This provided each surgeon's "policy" of cue use represented by the following equation:

$$Y_s = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m + e,$$

where Y_s = the surgeon's actual judgment of patient priority; X_m = the m cues on which the judgment is based; β_m = the m weights, which indicate the relative importance of each cue for the subject's judgment; and e = the random error term.

The regression model does not represent the exact cognitive processes that the surgeon goes through in making a judgment but is an attempt to describe the judgment in terms of the use of cues.[9]

Lens Model Statistics

To achieve the second aim of determining the cause(s) of lack of agreement, we needed to calculate lens model statistics. For every judgment (Y) made, each surgeon's regression model can then be used to generate a predicted judgment ($\hat{Y}$) using the cues. The coefficient of determination produced by the multiple regression is a measure of the multiple correlation of the actual judgments with the predicted judgments.[10] This value, designated R_S, represents each surgeon's cognitive control in using the equation (i.e., the consistency with which the surgeon's judgment model is applied over different patient vignettes). The correlation between different surgeons' predicted scores generates the value G (i.e.,

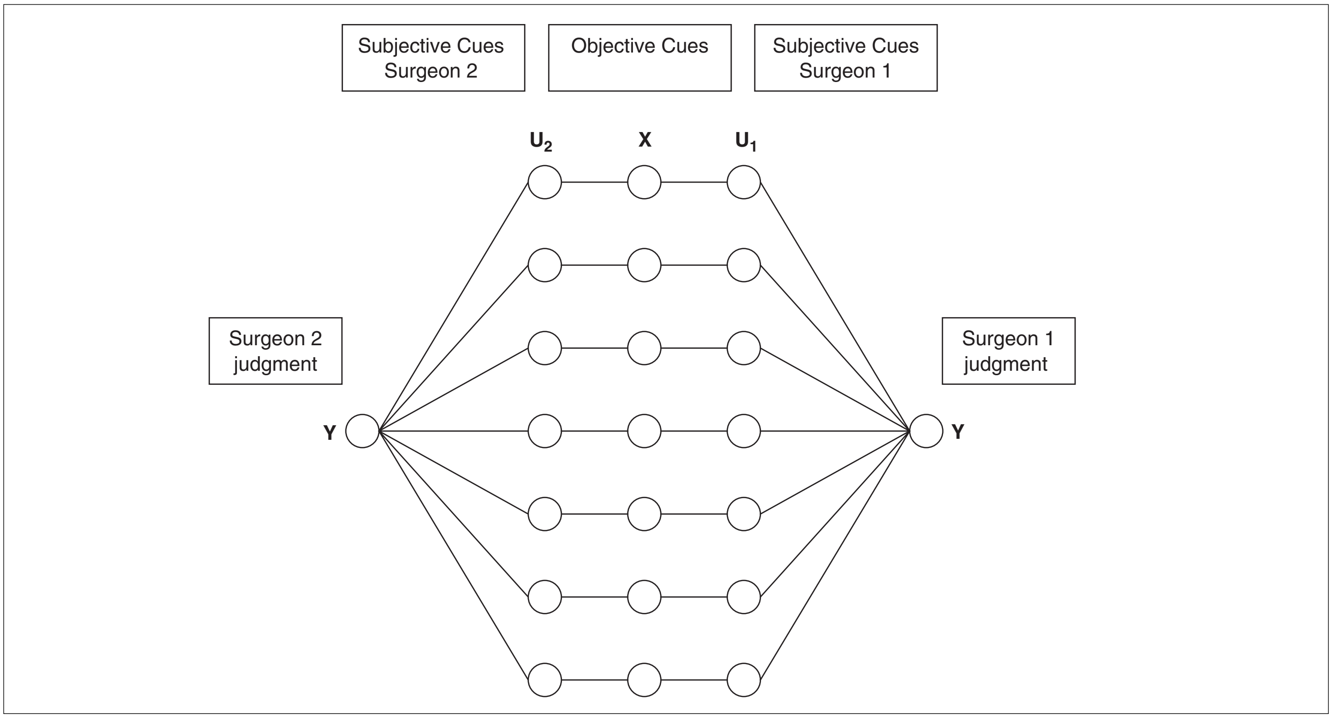

Figure 1 Modified Brunswik lens model (adapted from Stewart and Lusk[8]).

the similarity between surgeons' models).[10] The variable r_a is the correlation between different surgeons' actual scores (i.e., the agreement between surgeons). The correlation between the error terms produces the statistic C (i.e., a measure of any unmodeled component adding to agreement). The values R_S, G, r_a, and C are called *lens model statistics*.

The lens model statistics are represented by an equation.[8,11,12] In this situation when comparing surgeons, the lens model equation is

$$r_a = G\,R_{S1}\,R_{S2} + C\sqrt{(1 - R_{S1}^2)}\sqrt{(1 - R_{S2}^2)},$$

where r_a = the agreement between surgeons, G = the similarity in the policy models of the 2 surgeons, R_s = the cognitive control of the surgeon (1 or 2) over the model, and $C\sqrt{(1 - R_{S1}^2)}\sqrt{(1 - R_{S2}^2)}$ = the unmodeled component of the agreement.

The use of the objective and subjective cues enabled us to deconstruct R_S further.[8] R_S is represented as $V_{U.X.}$ multiplied by $R_{Y.U.}$. Both values ($V_{U.X.}$ and $R_{Y.U.}$) have a maximum of 1 and minimum of 0. $V_{U.X.}$ represents the reliability of information acquisition that occurs in taking the objective cue level in the vignette and turning it into a subjective cue level in the form of a score using the 50-mm VAS. A value of unity reflects perfect information acquisition; a value of 0 reflects total loss of information. It is calculated as a ratio of $R_{Y.X.}/R_{YU.}$. Similarly, $R_{Y.U.}$ represents the reliability of the information processing (or cognitive control of the surgeon), with a value of 1 representing perfect processing.

$R_{Y.X.}$ and $R_{Y.U.}$ are the coefficients of determination of a multiple regression using different independent variables. $R_{Y.X.}$ is calculated using the objective cues as independent variables, whereas $R_{Y.U.}$ is calculated using the subjective cues as independent variables. The dependent variable for both multiple regressions is the surgeon's judgment. The ability to deconstruct R_S into these values helps to determine the factors influencing the reliability and therefore identify problem areas that need to be remedied.

Case Vignettes

Thirty-two patient vignettes were constructed using a fractional factorial design on the 7 cues.[13] The format of the vignette followed that of demographics, presenting complaint, the history of the

ID	Demographic	Patient Complaint	History of Complaint	Past Medical History	Exam/Investigation
32	70 year old male	Inguinal hernia	Three year history of increasing size of R indirect hernia. Constant dragging pain. Pain exacerbated by movement. Intermittent bowel obstruction. Presented to Accident and Emergency on one occasion and forcibly reduced. Difficulty performing daily activities. Lives alone. Concerned about bowel symptoms and inability to perform self care.	Hypertension, CHF, IHD	Incarcerated Right indirect hernia, extends into scrotum.

Figure 2 Example patient vignette. Surgeons were asked to read the vignette and then prioritize the case using the 8-scale visual analog scale (VAS) (see Figure 4). CHF, congestive heart failure; IHD, ischemic heart disease.

presenting complaint, past medical history, and finally examination and investigation findings. The objective cues in the vignettes were determined to be either a high or low level (e.g., the chance or severity of future complications symptoms as being high or low). Verbal descriptions that matched these levels were then constructed and embedded in a vignette. For example, in the hernia vignettes, a high level for the future complications cue was that of having symptoms of impending bowel obstruction, and a low level was the absence thereof. These cue levels provided the objective cues in the model. The exception was that of diagnosis, where there were 4 categories—namely, inguinal hernias, hemorrhoids, cholelithiasis, and right-sided colon cancer. Surgeons vetted the cases for clinical realism.[6,10] An example vignette is included as Figure 2, and the cue descriptors are presented in an extracted format in Figure 3.

Subjects and Their Task

The subjects were 60 general surgeons randomly selected from those available on the New Zealand Medical Council Vocational Register. The surgeons were all familiar with the task of prioritizing patients for general surgery as part of their everyday clinical practice. Each surgeon rated all 32 vignettes. They rated the level of the 7 objective cues, using the 50-mm VAS. This interpretation of the cue level provided the subjective cues. The surgeons also made a global assessment of the patient's priority for surgery using a 100-mm VAS for the vignette (Figure 4). A score of 0 indicated that the patient was not indicated for elective surgery, whereas a score of 100 indicated that the patient should have his or her procedure on the next available operating list. Surgeons made all ratings in light of patients seen in clinical practice. None of the surgeons had used the 8-scale VAS priority-scoring tool previously. However, participants who worked at Auckland Hospital would have previously used a single VAS for the task of assessing global priority.

Assumption Testing

A number of assumptions are made when performing multiple regression analysis.[14] Unless an assumption (normality, linearity, autocorrelation, and homoscedasticity) was violated, we report the result without noting the performance on the assumption.

Analysis

Lens model statistics for each surgeon were calculated on the correlation of all pairs of judges. This was in contradistinction to calculating statistics for the surgeons against a gold standard. We undertook the former approach because no such gold standard existed, and we did not want to introduce an arbitrary standard. Due to the large number of correlations this produces, a 60 × 60 table for each statistic, these were summarized by transforming correlations using Fisher's z transformation and then calculating averages.[15]

Cue Title	Objective Cue Descriptor
Diagnosis	Inguinal Hernia
Treatment	Presented to Accident and Emergency on one occasion and forcibly reduced
Symptomatology	Three years of increasing size. Constant dragging pain. Pain exacerbated by movement.
Future Complications	Intermittent bowel obstruction. Incarcerated on examination.
Quality of Life	Difficulty performing daily activities. Lives alone.
Psychological Impact	Concerned about bowel symptoms and inability to perform self care.
Patient Characteristics	70 year old male. Past Medical History of Hypertension, CHF, IHD

Figure 3 *Extraction of the objective cues present in the vignette. CHF, congestive heart failure; IHD, ischemic heart disease.*

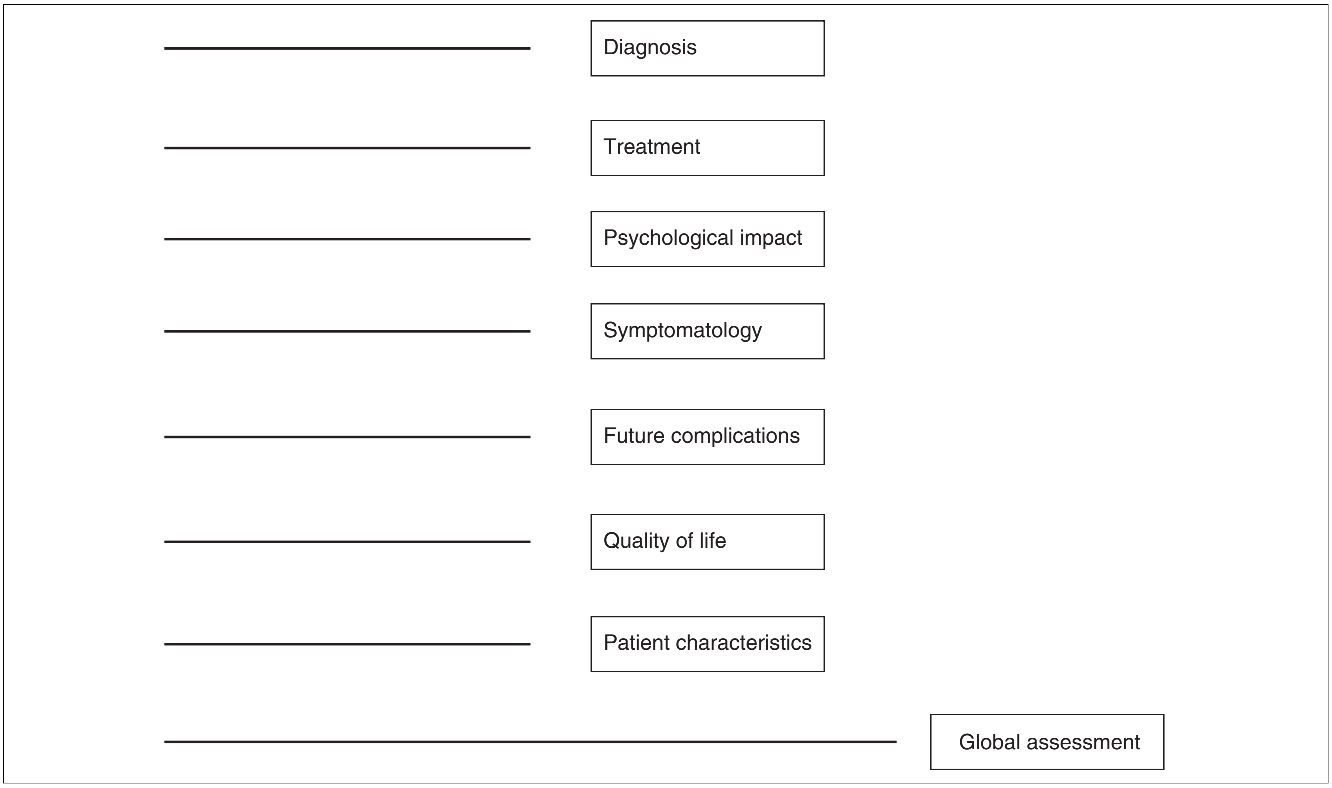

Figure 4 *Eight-scale visual analog scale (VAS) tool. For each vignette, surgeons were asked to indicate the level of the objective cue on the seven 50-mm VAS (thus giving the subjective cue level) and their judgment of patient priority on the 100-mm global assessment VAS.*

Cluster Analysis

The third aim was investigated using cluster analysis. Individual surgeons' policies underwent cluster analysis in an attempt to reveal groupings of surgeons with similar usage of cues. Cluster analysis was performed on the basis of the policies' standardized beta weights using Ward's method. The number of policy clusters was determined using visual inspection of the hierarchical tree diagram and an inverse scree plot of the number of clusters versus the semipartial R^2.

"Simple" Multiple Regression

Finally, to achieve the fourth aim of the study, multiple regression for the pooled priority judgments of the all the surgeons was undertaken. This approach gave us an overall indication of the predictability of the 8-scale VAS as a prioritization tool, which was not achieved with the judgment analysis approach of analysis of individual surgeons.

Statistical analysis in all cases was performed using SAS 8.02 (SAS Institute, Cary, NC).

RESULTS

Descriptive Statistics

The combined surgeons' mean priority score was 57. Individual surgeons' mean priority range was 23 to 77. The standard deviation of the combined surgeons was 28. The combined surgeons ranged over the entire possible scores of 0 to 100. Inspecting individual surgeons, the minimum scores ranged from 0 to 44, and the maximum scores ranged from 78 to 100.

Individual Policies

The individual surgeons' policies were determined using multiple regression techniques. We calculated the standardized cue weights, adjusted R^2, and P values for the multiple regression equation of each judge's policy. The adjusted R^2 had a mean of 0.77 (range, –0.02 to 0.96). Two subjects had nonsignificant policy models and were excluded from further analysis. This left 58 surgeons.

Standardized weights were chosen to represent surgeons' policies because of the difference in the means and standard deviations of the measurement units of both the cues and judgment measures.[10] The standardization process accounts for the different measures. Diagnosis was the highest weighted criteria in the most instances ($n = 17$). This was followed by quality of life (16), treatment (13), complications (13), psychological impact (1), symptomatology (1), and patient characteristics (0). In comparison, the least weighted criterion was patient characteristics (16), followed by symptomatology (13), psychological impact (12), treatment (7), quality of life (7), diagnosis (4), and complications (1). Numbers do not sum to 58 due to some criteria being equally weighted. The cluster analysis further helps with data reduction to understand any groupings of criteria weightings.

Table 1 shows the averages of all pairwise lens model statistics for each surgeon. The low agreement between surgeons is indicated by the average r_a for all surgeons being 0.48. The other statistics give an indication as to what contributes to this lack of agreement between surgeons. The low G statistic (0.47) indicates the lack of correlation between surgeons' policy models. In other words, surgeons disagree in how they use the cues and thus exhibit practice variation in their priority rankings. The value of the C statistic indicates low correlation between unmodeled policies. This indicates that there is unlikely to be another factor that all surgeons are using in determining their priority scores. The relatively high $V_{U.X.}$ shows that the lack of agreement is not due to poor subjective interpretation of the cues present in the vignettes. This also is evidence that the definitions of the cues were readily understood by the surgeons. The high $R_{Y.U.}$ indicates that there is not a problem with the cognitive control of the majority of surgeons in applying their policies.

Cluster Analysis

Inspection of the hierarchical tree diagram and the inverse scree plot revealed 2 groups who used the cues in different ways. One group predominantly used 3 cues and the other group 6 (Table 2). The groups consisted of 41 members and 17 members, respectively. These groups represent surgeons who have different philosophies as to the prioritization of patients for surgery.

Combined Surgeons

The multiple regression for the combined surgeons was undertaken to determine the predictability of

Table 1 Surgeons' Lens Model Statistics (Average of Comparisons with 57 Other Surgeons)

Surgeon	Average r_a	Average G	Average C	$V_{U.X.}$	$R_{Y.U.}$
1	0.58	0.60	0.00	1.00	0.86
2	0.30	0.24	0.06	0.81	0.89
3	0.34	0.40	0.02	0.91	0.87
4	0.50	0.53	0.02	0.96	0.90
5	0.33	0.37	–0.01	0.68	0.95
6	0.42	0.39	0.05	0.62	0.81
7	0.53	0.46	0.07	0.89	0.91
8	0.35	0.35	0.00	0.71	0.94
9	0.62	0.64	–0.02	0.95	0.97
10	0.53	0.52	0.01	0.84	0.98
11	0.51	0.52	0.07	0.91	0.84
12	0.58	0.52	–0.01	0.89	0.89
13	0.52	0.54	0.03	0.87	0.90
14	0.31	0.33	0.01	0.68	0.96
15	0.44	0.44	0.05	0.79	0.90
16	0.50	0.39	0.04	0.89	0.85
17	0.55	0.56	0.05	0.84	0.87
18	0.49	0.52	0.00	0.93	0.94
19	0.50	0.54	0.02	0.85	0.94
20	0.48	0.46	0.06	0.87	0.89
21	0.59	0.63	0.06	0.90	0.97
22	0.57	0.55	0.03	0.93	0.97
23	0.52	0.55	0.07	0.99	0.83
24	0.57	0.58	0.02	0.95	0.92
25	0.51	0.53	0.03	0.89	0.96
26	0.46	0.44	0.07	0.69	0.94
27	0.50	0.47	0.02	0.78	0.98
28	0.09	–0.05	0.10	0.59	0.93
29	0.34	0.32	0.05	0.67	0.91
30	0.54	0.55	–0.01	0.85	0.93
31	0.52	0.51	0.06	0.87	0.97
32	0.52	0.50	0.06	0.91	0.90
33	0.37	0.12	0.12	1.00	0.45
34	0.53	0.36	0.05	1.00	0.76
35	0.55	0.58	0.02	0.84	0.95
36	0.35	0.30	0.04	0.82	0.87
37	0.59	0.57	0.00	0.93	0.96
38	0.52	0.60	0.03	0.88	0.88
39	0.55	0.55	0.01	0.94	0.97
40	0.28	0.10	0.04	0.69	0.81
41	0.40	0.39	0.02	0.67	0.82
42	0.46	0.51	–0.01	0.95	0.92
43	0.29	0.26	0.06	0.57	0.96
44	0.57	0.56	0.08	0.97	0.98
45	0.55	0.58	0.02	0.90	0.89
46	0.57	0.52	0.04	0.82	0.96
47	0.52	0.55	–0.04	0.88	0.95
48	0.48	0.50	0.01	0.76	0.96
49	0.60	0.59	0.05	1.00	0.97
50	0.60	0.60	0.07	0.94	0.90
51	0.48	0.46	–0.04	0.83	0.93
52	0.32	0.36	–0.07	0.76	0.89
53	0.54	0.53	0.02	0.83	0.93
54	0.46	0.50	0.01	0.88	0.96
55	0.63	0.74	0.01	0.92	0.94
56	0.51	0.53	0.07	0.94	0.76
57	0.27	0.21	0.05	0.67	0.94
58	0.49	0.58	–0.01	0.73	0.83
59	0.42	0.40	–0.01	0.55	0.98
60	0.40	0.22	0.10	1.00	0.61
Average (for all surgeons)	0.48	0.47	0.03	0.84	0.91

r_a = the agreement between surgeons; G = the similarity in the policy models of the surgeons; R_s = the cognitive control of the surgeon over his or her model; C = the unmodeled component of agreement; $V_{U.X.}$ = reliability of information acquisition; $R_{Y.U.}$ = reliability of information processing.

the 8-scale VAS. It resulted in a coefficient of determination (R^2) of 0.66 with 6 cues reaching significance (Table 2). Coefficients such as this can be characterized as slight (0–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), and almost perfect (0.81–1.00).[16] When applied to our result, they indicate that the use of the 8-scale VAS results in "substantial" predictability of the generated priority score.

DISCUSSION

This study intended to describe quantitatively surgeons' judgment of patient priority for elective general surgery. We were interested in how surgeons used the criteria that they had previously indicated they believed they used.[6] We were also looking to identify the cause(s) of the hypothesized lack of agreement between surgeons in the undertaking of this task.

Multiple Regression

Surgeons' use of the criteria proved to be conducive to modeling in all but 2 instances. The cue use of these 2 surgeons was not able to be captured at a significant level with 32 vignettes. This number of vignettes is less than the optimum of 35, which is based on 5 vignettes per cue. With more vignettes, their judgment may have been able to be more fully analyzed. However, we had to take into account the cognitive load on the surgeons, considering that this study was performed in conjunction with a reliability test in which 2 other potential priority-scoring tools were being tested. The remainder of the surgeons showed relatively high cognitive control, indicating that the surgeons were consistent in the application of their policy models.

Table 2 Surgeon Policy Clusters

	Cluster A (n = 41), R^2 = 0.73 (<0.0001)		Cluster B (n = 17), R^2 = 0.55 (<0.0001)		Combined Surgeons (n = 58), R^2 = 0.66 (<0.0001)	
	Weight	P Value	Weight	P Value	Weight	P Value
Diagnosis	0.19	<0.0001	–0.02	0.6694	0.13	<0.0001
Treatment	0.11	<0.0001	0.48	<0.0001	0.21	<0.0001
Symptoms	0.06	0.0005	0.06	0.1384	0.07	<0.0001
Complications	0.33	<0.0001	0.30	<0.0001	0.32	<0.0001
Quality of life	0.25	<0.0001	0.11	0.0494	0.22	<0.0001
Psychological	0.01	0.5295	–0.01	0.7642	0.01	0.8556
Patient characteristics	0.06	0.0015	0.01	0.8459	0.05	0.0057

Agreement

Our interest lay in the poor agreement (r_a = 0.48) between surgeons in their judgment of priority. Poor agreement in general leads to concerns that if the same patient were to be prioritized by different surgeons, he or she might receive different priority scores, thereby affecting the patient's access to surgery. Thus, access to elective surgery may in fact be determined by the surgeon to whom the patient is referred.

The lens model equation allowed the deconstruction of this value. Due to its high value, information processing ($R_{Y.U.}$) did not negatively influence agreement. In addition, the information acquisition ($V_{U.X.}$) was gratifyingly high and therefore did not contribute to the poor agreement. The influence of unmodeled knowledge (C) was also minimal.

Our major finding was that the cause of poor agreement between surgeons mostly lay in the poor correlation (G) between how individual surgeons use the criteria. Variation in weighting applied to criteria may be the result of the lack of formal training in the setting of priority for surgery. Furthermore, it may be due to the sporadic way in which priority setting was implemented through the country.[2] It may also be contingent on the areas of surgical specialization or the institutional biases of the participant surgeons. We investigated the possibility that it is the result of different philosophical or ethical viewpoints in the cluster analysis. Formal study of the reasons why surgeons apply certain weights may be a topic for further investigation.

The different weighting that surgeons place on the criteria is a form of individual practice variation in priority setting. Methods should be sought to decrease this variation. One such method, known as cognitive feedback (CFB), has been postulated to be helpful with this form of variation.[8,17]

CFB, in 1 form, uses the beta weights of each cue to enable the development of consensus between surgeons. In contrast, conventional outcome feedback, in this setting, would only feedback the priority score. CFB has been shown to be more successful in developing consensus than conventional outcome feedback.[18] This is only one of several possible interventions. However, we propose future studies to test this hypothesis in our setting.

Assumption Testing

The results of the assumption testing supported the use of a multiple regression model. Tests were conducted at the P = 0.01 level.

Clusters

The cluster analysis enabled us to identify 2 groupings of surgeons who use the criteria in a different manner. One group takes more account of future complications, quality of life, and the diagnosis, whereas the other group considers the treatment and future complications. However, it should be noted that symptoms are weighted the same in both groups, and the lack of significance in cluster B is probably related to sample size. This is indicative of at least 1 of the limitations with cluster analysis—namely, development of spurious groupings. That is, there may not be a useful way to distinguish the data points, but the analysis forces it to occur.[19] Furthermore, different methods of cluster analysis may result in different groupings.[20] However, in this example, even should symptoms be included in both clusters, the remaining criteria are not likely to reach significance.

These 2 distinct groups of surgeons may reflect these differences as "philosophies" of the prioritization of patients for surgery. They could result from different ethical standpoints regarding how to prioritize. For example, 1 group may consider the urgency of the need, as reflected by the use of quality of life, whereas others may think of the capacity to benefit from the surgery, as indicated by the use of the treatment cue. These different policy groupings may lead to a different order of patient priority and therefore result in different people receiving surgery. However, it must be emphasized that these groupings are descriptive and not normative. These clusters, as descriptive groupings, provide an initial point for CFB and the subsequent development of consensus on criteria use. Nevertheless, cluster analysis remains an exploratory tool for investigating the data, and the limitations mentioned above must be borne in mind.

Combined Surgeons

The combined regression enabled us to determine the predictability of 8 visual analog scales as a tool to assign priority to patients. The R^2 indicates that the use of the 8-scale VAS as a tool for prioritization results in "substantial" ability to predict the clinical judgment of priority. If the levels of the 7 criteria were inputted, the weightings and error term identified by this regression would thus predict the correct clinical judgment of priority 66% of the time, if used as a "formula" to generate the priority score. Although this is gratifying, it needs to be reexamined by cross-validating the model using real patients in a clinical setting. This would be important to dissuade surgeons of their skepticism as to the validity of a priori weighted scoring tools in which they have not had conscious input.

Other Work

Much work is being conducted on the booking system in New Zealand and similar waiting list systems in the United Kingdom and Canada. New Zealand is conducting a large research project evaluating clinical priority assessment criteria. Currently, this involves undertaking a conventional test-retest study to determine the inter- and intrasurgeon reliability of a number of priority scoring tools over 3 surgical specialties. Canada has undertaken a program to determine appropriate criteria and scoring tools.[21] The United Kingdom has not developed national tools but has opted for a process engineering approach.[22] However, judgment analysis has been used to investigate priority setting in cardiac surgery.[23,24]

No previous studies have been undertaken that investigate the reliability of priority scoring tools used in New Zealand. One study[25] has been undertaken that compared the scores assigned by different priority-scoring tools. It showed that priority-scoring tools with arbitrary a priori weighting of criteria did not reflect clinical judgment. This resulted in the rejection of such arbitrary tools by clinicians. Nevertheless, despite the lack of good supportive evidence supporting their use, priority-scoring tools such has those used in New Zealand have been taken up and modified by different UK health trusts.

Summary

This project has taken a descriptive approach to how surgeons prioritize patients for elective general surgery. This has been primarily because of the paucity of evidence for criteria on which to prioritize elective general surgical patients.[26] Although a normative approach to prioritization (with a priori determined weightings) may in theory lead to greater agreement between surgeons, the lack of reflection of clinical judgment has shown this to be unworkable.[3,25]

We have found significant lack of agreement. However, the main cause for this is the lack of consensus between surgeons in the weighting of cues. The cluster analysis shows that these may be the result of 2 different philosophies of prioritization based around ethical principles. Cognitive feedback may provide a method to reduce the variation by enabling discussion and the development of consensus on cue weighting.

The contribution of this study is the development of a priority-scoring tool that enables clinical judgment to be used in the priority-setting process. Furthermore, the clinical judgment can occur in a transparent fashion open to inspection and critique. It also helps to identify ways in which we may be able to improve iteratively the reliability of the priority-scoring tools. Therefore, the use of a priority-scoring tool based on judgment analysis enables us to meet the different requirements held by clinicians (reflecting clinical judgment) and the Ministry of Health (transparency).

These results need to be cross-validated in a clinical setting. Furthermore, the VAS tool needs to undergo tests of reliability and predictive validity to investigate the influence on patient outcome.

REFERENCES

1. Fraser G, Alley P, Morris R. Waiting Lists and Waiting Times: Their Nature and Management. Wellington: National Advisory Committee in Core Health and Disability Support Services; 1993.

2. Gauld R, Derrett S. Solving the surgical waiting list problem? New Zealand's 'Booking System.' Int J Health Plann Manage. 2000;15:259–72.

3. Dennett E, Parry B. Generic surgical priority criteria scoring system: the clinical reality. N Z Med J. 1998;111:163–6.

4. MacCormick A, Plank L, Robinson E, Parry B. Prioritizing patients for elective surgery: clinical judgement summarized by a linear analogue scale. Aust N Z J Surg. 2002;72:613–7.

5. Hammond K. Probabilistic functioning and the clinical method. Psychol Rev. 1955;62:255–62.

6. MacCormick A, Parry B. Identification of criteria for the prioritisation of patients for elective general surgery. J Health Serv Res Policy. 2004;9:28–33.

7. Brehmer B. Effects of communication and feedback on cognitive conflict. Scan J Psychol. 1971;12:205–16.

8. Stewart T, Lusk C. Seven components of judgmental forecasting skill: implications for research and the improvement of forecasts. J Forecasting. 1994;13:579–99.

9. Hoffman P. The paramorphic representation of clinical judgment. Psychol Bull. 1960;57:116–31.

10. Cooksey R. Judgment Analysis: Theory, Methods, and Applications. San Diego: Academic Press; 1996.

11. Hursch C, Hammond K, Hursch J. Some methodological considerations in multiple-cue probability studies. Psychol Rev. 1964;71:42–60.

12. Tucker L. A suggested alternative formulation in the development by Hursch, Hammond, and Hursch and by Hammond, Hursch, and Todd. Psychol Rev. 1964;71:528–30.

13. Cochran W, Cox G. Experimental Designs. 2nd ed. New York: John Wiley; 1957.

14. Katz M. Multivariable Analysis: A Practical Guide for Clinicians. Cambridge (UK): Cambridge University Press; 1999.

15. Snedecor G, Cochran W. Statistical Methods. 7th ed. Ames: Iowa State University Press; 1980.

16. Landis R, Koch G. The measurement of observer agreement for categorical data. Biometrics. 1977;33:59–74.

17. Doherty M, Balzer W. Cognitive feedback. In: Brehmer B, Joyce C, eds. Human Judgment: The SJT View. Amsterdam: Elsevier Science; 1988.

18. Klayman J. On the how and why (not) of learning from outcomes. In: Brehmer B, Joyce C, eds. Human Judgment: The SJT View. Amsterdam: Elsevier Science; 1988.

19. Manly B. Multivariate Statistical Methods. London: Chapman & Hall; 1988.

20. Chatfield C, Collins A. Introduction to Multivariate Analysis. London: Chapman & Hall; 1980.

21. McDonald P, Shortt S, Sanmartin C, Barer M, Lewis S, Sheps S. Waiting Lists and Waiting Times for Health Care in Canada: More Management!! More Money?? Ottawa, Canada: Health Canada; 1998.

22. Kipping R, Meredith P, McLeod H, Ham C. Booking Patients for Hospital Care: A Progress Report. Birmingham (UK): Health Services Management Centre, University of Birmingham; 2000.

23. Kee F, McDonald P, Kirwan JR, Patterson CC, Love AHG. The stated and tacit impact of demographic and lifestyle factors on prioritisation decisions for cardiac surgery. QJM. 1997;90:117–23.

24. Kee F, McDonald P, Kirwan JR, Patterson CC, Love AHG. Urgency and priority for cardiac surgery: a clinical judgment analysis. Br Med J. 1998;316:925–9.

25. Dennett E, Kipping R, Parry B, Windsor J. Priority access criteria for elective cholecystectomy: a comparison of three scoring methods. N Z Med J 1998;111:231–3.

26. MacCormick A, Collecutt W, Parry B. Prioritising patients for elective surgery: a systematic review. ANZ J Surg. 2003;73:633–42.