ELSEVIER

Health Policy 37 (1996) 139–152

HEALTH policy

Panellist consistency in the assessment of medical appropriateness

Joseph McDonnell[a,b,*], Annejet Meijler[a,b], James P. Kahan[c], Steven J. Bernstein[d,e], Henk Rigter[a,b]

[a]*Institute for Medical Technology Assessment, Erasmus University, P.O. Box 1738, 3000 DR, Rotterdam, The Netherlands*
[b]*Inst. of Social Medicine, Erasmus University, Rotterdam, The Netherlands*
[c]*RAND/European-American Center for Policy Analysis,Delft, The Netherlands*
[d]*University of Michigan, Ann Arbor, Michigan, USA*
[e]*Veterans Administration Health Services Research Field Program, Ann Arbor, Michigan, USA*

Received 27 February 1996; accepted 4 March 1996

Abstract

Where information about the appropriateness of a surgical procedure is lacking, expert panels have been used to establish guidelines for medical practitioners. Such a panel was convened to assess the appropriateness of percutaneous transluminal coronary angioplasty and coronary artery bypass graft surgery in the Netherlands. The panel, consisting of interventional cardiologists and cardiothoracic surgeons, used a modified Delphi process to rate 1126 clinical indications over two rounds. This article describes the degree of change in both agreement amongst members and in the appropriateness ratings over the two rounds, and examines the internal consistency of the ratings of individual panellists. Over the two rounds, agreement increased. Although most appropriateness ratings remained unchanged, there was significant movement from equivocal ratings to determinate ratings. While individual members showed some degree of inconsistency in their scoring, the panel as a whole scored very consistently. The observed changes in appropriateness were consistent with expectations, showing that the appropriateness method is used logically and consistently by panellists.

Keywords: Modified Delphi process; Agreement; Appropriateness

* Corresponding author. Tel.: +31 10 4088571; fax: +31 10 4522511.

PII S0168-8510(96)00828-7

1. Introduction

In the absence of large clinical trials, there may be considerable uncertainty over the appropriateness of a given medical treatment in relation to specific symptoms and the results of tests carried out on a patient. Several studies have augmented the scientific evidence with a panel of experts to establish guidelines for medical practice [1–3]. In 1991, the Dutch Inventory of Invasive Coronary Atherosclerosis Treatments (DUCAT) group assembled a panel of experts to establish a rulebase for the use of coronary artery bypass graft surgery (CABG), percutaneous transluminal coronary angioplasty (PTCA) and medical therapy (MED) for patients with coronary artery disease [4,5]. The rulebase was developed using a modified Delphi process; panel members were asked to give their opinion as to the appropriateness of the invasive procedures on two occasions, once at home with no contact between members, and later after a meeting of all members where the results of the first round were discussed. The modified Delphi process allows individual members to review and possibly modify their opinions through exposure to the opinions of other members and discussion of the differences in opinion.

This process permits a panel member to show substantial changes in opinion. There are essentially three reasons why such changes might occur:

(1) Some members of a given panel will not be familiar with their task and may make errors in recording their opinions (for example, scoring a procedure as 'appropriate' where 'inappropriate' was intended). Such errors may occur in either round, though they would be more likely to occur in the first round when panel members are less familiar with the process. However, in a multi-round process, factors such as mental fatigue would occur in later chapters of each round.
(2) The re-rating process may cause a panellist to reappraise his own reasons for his original decision; such a review is independent of other panel members.
(3) The feedback of others' ratings and discussion may cause some individual panel members to modify some of their opinions.

Whatever the reason, some panel members will modify their personal opinion as to the appropriateness of an invasive procedure; this expected change is an essential characteristic of the modified Delphi process. To gain insight into the presence and effect of the factors leading to change, we examined the ratings produced by DUCAT panel members. Comparison of the results of the two rounds allows assessment of the consistency of the panel and the individual members.

2. Methods

The DUCAT panel consisted of six interventional cardiologists and six cardiothoracic surgeons, one from each of the 12 of the revascularisation centres then operating in the Netherlands. Both academic and non-academic practitioners were

included among the 12 panellists. The initial list of indications for PTCA and CABG was developed in the United States by a multi-specialty panel of cardiologists, cardiothoracic surgeons and internists after a comprehensive literature review [6,7]. Following consultation with the members of the DUCAT panel, this list was modified to suit the requirements of the DUCAT study and to reflect clinical thinking and practice in the Netherlands. Each indication corresponded to a clinical scenario consisting of a combination of clinical and laboratory factors considered important in the decision to recommend revascularisation and was defined so that patients with a given indication could be considered to be reasonably homogenous with regards to each treatment. The indications were divided into eight major clinical categories corresponding to the patient's primary underlying medical condition (Table 1 lists these groups together with the number of indications within each group). An example of an indication is 'A patient has severe angina while on optimal medical therapy, is at low operative risk, has an ejection fraction of more than 20% and coronary angiography has demonstrated significant stenosis in the proximal left anterior descending artery with a type C lesion'. Each term in the indication was defined in a glossary that accompanied the indications.

For each indication in clinical categories 1–6 of Table 1, panel members made three two-way comparisons: (1) the appropriateness of PTCA compared to medical therapy (PTCA-MED); (2) the appropriateness of CABG compared to medical therapy (CABG-MED); and (3) the appropriateness of CABG compared to PTCA (CABG-PTCA) on a scale of 1–9. For the comparisons PTCA-MED and CABG-MED, an indication for which the revascularisation was extremely appropriate was rated as 9, with less appropriate indications receiving progressively lower scores down to 1 (revascularisation extremely inappropriate); for the comparison CABG-PTCA, a score of 9 indicated an extreme preference for CABG, a score of 1 an extreme preference for PTCA. An invasive treatment was defined to be appropriate if the expected health benefit exceeded the expected negative consequences; similarly, PTCA was preferred to CABG if the PTCA would benefit the patient more than CABG, given the risks involved. For Groups 7 and 8, only one comparison per patient was possible (CABG vs. MED and PTCA vs. MED, respectively). Panel

Table 1
Clinical characteristics of indications

Group	Description (No. of indications)
1	Asymptomatic (96)
2	Chronic stable angina pectoris (480)
3	Unstable angina pectoris (144)
4	Acute myocardial infarction (68)
5	Post myocardial infarction (200)
6	Near sudden death (96)
7	CABG combined with valve surgery (2)
8	Palliative PTCA (40)

members were explicitly instructed not to include considerations of cost factors in their deliberations.

The panel completed their ratings twice, once without the opportunity of discussion with fellow members of the panel and again, one and a half months later, during a 2-day meeting in August 1991 with other panel members. Before the second round, each member received a summary of the results of the first round giving, for each indication, his own score together with the distribution of the scores of all panel members; in this way, individual ratings remained anonymous. The discussion during the meeting was informal and members were encouraged to voice their opinions. Although the discussion centred on those indications where there were large differences of opinion, the course of the discussion was not limited to choosing a rating number. Instead, the list of indications and the definitions of some of the variables were also refined. The panel felt that some variables had been too finely detailed, whereas others needed further refinement. For example, the ratings for patients with normal left ventricular function, mild left ventricular dysfunction and moderate left ventricular dysfunction were so similar that these three categories could be taken together. On the other hand, for post myocardial infarction patients with continuing or recurrent pain, the panel felt that it was necessary to make separate comparisons for patients receiving optimal medical therapy and those receiving less than optimal therapy. Therefore, comparison of the results in the two rounds is done only on indications common to both rounds (for the post myocardial infarction patients, indications involving optimal medical therapy in round 2 were used for the comparison, as the physicians assumed in round 1 that the therapy offered was optimal). The analysis is further restricted to Groups 1–6 since in Groups 7 and 8, a three-way comparison was not applicable.

The scores were analysed with respect to the degree of agreement among panel members and appropriateness. The definitions used were similar in nature to those used by RAND in previous studies of the appropriateness of revascularisation [6–8]. The panel were said to 'agree' if at least nine of the 12 scores lay in one of the three-point regions 1–3, 4–6, or 7–9, 'disagree' if at least four of the scores lay in each of the regions 1–3 and 7–9; otherwise, the panel were said to be 'indeterminate' with respect to agreement. An invasive treatment was said to be 'appropriate' if the median score lay in the region 7–9 without 'disagreement', 'inappropriate' if the median score lay in the region 1–3 without 'disagreement', and 'equivocal' if the median score lay in the region 4–6 or if there was disagreement (Table 2). Similarly, for the comparison CABG-PTCA, the panel's choice could be expressed as 'preference for CABG', 'preference for PTCA' or 'equivocal'.

With these data from two rounds, we examine four questions which address the issue of panel consistency:

(1) To what degree did 'agreement' change between the two rounds? (agreement would naturally be expected to increase from round 1 to round 2).
(2) To what degree did 'appropriateness' change?

Table 2
Panel appropriateness rating based on degree of agreement among panelists and median panel rating

Median panel score of appropriateness	Degree of agreement		
	Agreement	Indeterminate	Disagreement
1–3	Inappropriate	Inappropriate	Equivocal
4–6	Equivocal	Equivocal	Equivocal
7–9	Appropriate	Appropriate	Equivocal

See Section 2 for definitions.

(3) Do the individual ratings show internal (logical) consistency (as measured by the transitivity of the three ratings for each indication)?
(4) Can a learning process be discerned for some panel members?

3. Results

Although an increase in agreement was not explicitly sought, 'agreement' as defined should be greater in round 2 than in round 1 while 'disagreement' and 'indeterminate' can both be expected to decrease over rounds. This was found to be true for all three comparisons (Table 3). Roughly one-sixth of all ratings shifted from less-than-agreement to agreement, while roughly one-half of disagreements were resolved.

The overall pattern was evident in all six clinical categories.

Changes in the distribution of 'appropriateness' were also found. The results, for each of the two-way comparisons, are given in Tables 4–6.

Table 3
Agreement ratings per comparison per round

Comparison	Round	% Agreement	% Disagreement	% Indeterminate	*P*-value*
PTCA-MED	1	43.3	7.1	49.6	<0.001
	2	62.1	3.5	34.4	
CABG-MED	1	37.7	9.2	53.0	<0.001
	2	53.7	5.4	41.0	
CABG-PTCA	1	34.7	1.5	63.8	<0.001
	2	60.1	0.6	39.3	

*Test of marginal homogeneity.

Table 4
Appropriateness of PTCA compared to medical therapy by Rating round (number of indications as a percentage given in brackets)

Round 1	Round 2			
	Appropriate	Equivocal	Inappropriate	Total
Appropriate	175 (16.1)	1 (0.1)	0 (0.0)	176 (16.2)
Equivocal	47 (4.3)	215 (19.8)	44 (4.1)	306 (28.2)
Inappropriate	0 (0.0)	27 (2.5)	575 (53.0)	602 (55.5)
Total	222 (20.5)	243 (22.4)	619 (57.1)	1084

Table 5
Appropriateness of CABG compared to medical therapy by Rating round

Round 1	Round 2			
	Appropriate	Equivocal	Inappropriate	Total
Appropriate	336 (31.0)	3 (0.3)	0 (0.0)	339 (31.3))
Equivocal	54 (5.3)	278 (25.6)	40 (3.7)	372 (34.3)
Inappropriate	0 (0.0)	19 (1.8)	354 (32.7)	373 (34.4)
Total	390 (36.0)	300 (27.7)	394 (36.3)	1084

Table 6
The appropriateness of CABG compared to PTCA by Rating round

Round 1	Round 2			
	Appropriate	Equivocal	Inappropriate	Total
Appropriate	261 (24.1)	9 (0.8)	0 (0.0)	270 (24.9)
Equivocal	35 (3.2)	405 (37.4)	41 (3.8)	481 (44.4)
Inappropriate	0 (0.0)	6 (0.6)	327 (30.2)	333 (30.7)
Total	296 (27.3)	420 (38.7)	368 (33.9)	1084

The transition patterns for all three comparisons are generally similar. The number of comparisons which remained unchanged was high (89% PTCA-MED, 89% CABG-MED, 92% CABG-PTCA). Correspondingly, the median scores underlying the rating categorisation were stable; only about 2.5% of the ratings had a shift of median greater than 2 points on the 9-point scale. There was no indication that shifted two steps (inappropriate to appropriate or vice-versa) between rounds;

The transition patterns for all three comparisons are generally similar. The number of comparisons which remained unchanged was high (89% PTCA-MED, 89% CABG-MED, 92% CABG-PTCA). Correspondingly, the median scores underlying the rating categorisation were stable; only about 2.5% of the ratings had a shift of median greater than 2 points on the 9-point scale. There was no indication that shifted two steps (inappropriate to appropriate or vice-versa) between rounds; indeed, movement away from an appropriate or inappropriate rating was rare. There was a significant amount of movement ($P < 0.001$ for all three ratings) from an equivocal rating to an appropriate or inappropriate one, with that movement not being dominated by either an upward (towards appropriate) or downward (towards inappropriate) shift.

The vast majority of the shifts in appropriateness category were caused by the movement of the group median as opposed to the disappearance of disagreement; for example, of the 91 PTCA-MED shifts from equivocal, only seven were for resolution of disagreement without a change of median. Analyses not shown here indicate that all six clinical characteristic groups behaved similarly with respect to changes in appropriateness category.

3.1. Variations in individual scores

Although the panel group score is quite stable, panel members changed their individual scores, in some cases quite dramatically. For each member, we calculated the frequency and percentage of indications changed by three or more points from round 1 to round 2. We also calculated these statistics for 'severe' change, defined as a shift from appropriate (7–9) to inappropriate (1–3) or vice-versa. We reasoned that severe changes would most likely come from correction of a rating 'error' or from redefinition of a term, whereas 'moderate' changes of 3 or more points that were not severe would most likely reflect a true opinion change.

Table 7 shows how individual panellists changed. Overall, individual members changed between 114 (3.5%) and 583 (17.9%) of the more than 3200 ratings they made. Panellists could be grouped as making relatively frequent (309 or more) changes (members 1, 3, 4, 9, 10 and 12) or relatively rare (231 or fewer) changes (members 2, 5, 6, 7, 8 and 11). Among panellists, between 10 (0.3%) and 236 (7.2%) severe changes were made. Almost the same grouping of panellists could be made for changes as for severe changes, with five members having 89 or more and seven members having 69 or fewer. Panellist 12 was an outlier, although he was classified as a frequent changer, he made the fewest severe changes. For this panellist, only 3% of all changes were severe, while for the rest of the panellists, that statistic ranged from 18.6% to 40.5%.

We conclude that understanding of the process and definitions was a significant factor in panellist change propensity, but that this factor was not (except for panel member 12) related to general propensity to change ratings.

For each member, we calculated a measure of how that individual's scores differ from those of the group. For the ith member, this mean absolute deviation score is defined by $\mathrm{IMAD}_i = \Sigma_j |X_{ij} - M_j|/n$ where X_i = the score given by panel member

Table 7
Individual panelist ratings changes

Member	1	2	3	4	5	6	7	8	9	10	11	12
Number of ratings	3258	3222	3257	3253	3254	3257	3231	3241	3262	3253	3269	3252
Change ≥3 points	404	145	583	322	231	114	210	188	460	309	219	335
% of ratings with change ≥3 points	12.4	4.5	17.9	9.9	7.1	3.5	6.5	5.8	14.1	9.5	6.7	10.1
Severe changes[a]	116	27	236	89	66	38	45	63	111	98	69	10
% of all changes	28.7	18.6	40.5	27.6	28.6	33.3	21.4	33.5	24.1	31.6	31.5	3.0

[a]Severe change — a change from the appropriate range (7–9) to the inappropriate range (1–3) or vice versa.

i for indication j, M_j = the median panel score for indication j and n = the number of indications. This measure contrasts with the usual MAD which is a measure of the variation present in a group of observations.

Each member's IMAD was smaller in round 2 than in round 1, as would be expected in the presence of increasing agreement (Table 8).

Panel members 1, 3, 4, 9 and 12 had the greatest change in IMAD, indicating a tendency to score more towards the median score in round 2 than in round 1; their scores were generally more in line with those of other members of the panel in round 2 than they had been in round 1. Panellist 5 had the largest IMAD but the difference between the MADs was one of the smallest; his scores differed, on average, by approximately 2 points from the group median in both rounds. Members 1 and 3 differed much more from their colleagues in round 1 than in round 2, confirming that these members did indeed undergo a learning process. Although panellist 9 had a large number of substantial changes in his scores, in neither round did he differ markedly from his colleagues.

4. Transitivity

To check on the internal consistency of the ratings, we examined the transitivity of the panel appropriateness score for the two rounds. Here, we say $X > Y$ (is preferred to) if X is rated appropriate compared to Y (or Y is inappropriate compared to X). We consider $X = Y$ if the X vs. Y rating is equivocal. We then say that an indication is transitive if the following condition is fulfilled:if $X \geq Y$ **and** $Y \geq Z$ **then** $X \geq Z$; if one or more of the relationships X-Y, Y-Z is strict (preference = '>'), then $X > Z$. If an indication is not transitive, it is said to be intransitive.

Table 8
Individual mean absolute deviations, by round

Member	IMAD round 1	IMAD round 2	Difference between rounds
1	1.5678	0.9133	0.6545
2	1.1345	1.0132	0.1213
3	1.9500	1.2109	0.7391
4	1.3796	0.8915	0.4882
5	2.0816	1.9714	0.1102
6	0.9734	0.9194	0.0540
7	1.2711	1.2168	0.0543
8	1.6265	1.2764	0.3501
9	1.3073	0.8416	0.4657
10	1.5580	1.2057	0.3522
11	1.3562	1.0824	0.2738
12	1.3996	0.8998	0.4998
Mean over members	1.4671	1.202	0.3469

We make the distinction between strong and weak transitivity. An indication is strongly intransitive if $X > Y$ and $Y > Z$ but $Z \geq X$. An indication is weakly intransitive if $X = Y$ and $Y = Z$ but $X > Z$ (or $Z > X$). Note that a weakly intransitive indication may not represent a logical inconsistency. For example, the panel may have a very slight preference for PTCA over MED (but not sufficiently strong for PTCA to be rated as 'appropriate') and a slight preference for CABG over PTCA (again, not sufficiently strong for CABG to be judged as 'preferred' to PTCA); the combinations of these individually weak preferences may cause the panel to decide that CABG is indeed 'preferred to MED'. Such an indication would be rated as weakly intransitive under the above definition. We have not made the further distinction between 'possibly' transitive and other weakly intransitive indications.

Strongly intransitive indications do represent a logical inconsistency. The presence of a relatively large number of strongly intransitive indications would possibly imply that the three two-way comparisons were not being carried out on the same scale, i.e. one or more extra factors, such as the cost of the procedure, were entering into the panel's considerations. Clearly, the presence of many strongly intransitive indications would indicate that the panel was not able to clearly compare the three treatments.

Since each of the three comparisons has three possible outcomes (appropriate/equivocal/inappropriate or PTCA/equivocal/CABG), there are 27 possible combinations, of which 13 are transitive, eight are strongly intransitive, and six are weakly intransitive. These combinations are listed in Table 9, together with the type of intransitivity and the number of indications with that combination found in the panel data.

Taking panel median ratings, there were 10 (0.9%) strongly intransitive combinations in round 1 and 14 (1.3%) in round 2. There were 232 (21.4%) weakly intransitive combinations in round 1 and 186 (17.2%) in round 2; this reduction was significant ($P < 0.001$, McNemar's test). Overall, more than 80% of the indications are rated transitively by the panel.

For all indications in round 2 in which strong intransitivity was found, the indications involved a C lesion, while in round 1, seven of the indications involved a C lesion. Similarly, almost all the intransitive indications in round 2 involved two-vessel disease (with or without the involvement of the proximal left anterior descending) or one vessel disease with proximal left anterior descending involvement, and low operative risk; in round 1, several indications for left main disease or three-vessel disease, or for high operative risk were intransitively scored. The treatment of C lesions is an area where change is rapid and this is reflected in the fact that the panel opinion is not clearly expressed; similarly, the treatment preferred for patients suffering from moderate disease is an area where less clarity is to be found.

The fact that few of the indications were strongly intransitive indicates that the panel as a whole generally understood the method. Individual members however showed considerable variation in the number of intransitive indications (Table 10).

Table 9
Group patterns of transitivity

PTCA-MED	CABG-MED	CABG-PTCA	Type of intransitivity	Round 1	Round 2
P	C	C		0	3
P	C	E		21	34
P	C	P		28	49
P	E	C	Strong	0	0
P	E	E	Weak	12	2
P	E	P		94	104
P	M	C	Strong	0	0
P	M	E	Strong	0	0
P	M	P		21	30
E	C	C		28	35
E	C	E	Weak	56	37
E	C	P	Strong	0	0
E	E	C	Weak	2	3
E	E	E		85	41
E	E	P	Weak	39	23
E	M	C	Strong	0	0
E	M	E	Weak	18	17
E	M	P		78	87
M	C	C		198	218
M	C	E	Strong	8	14
M	C	P	Strong	0	0
M	E	C		41	34
M	E	E	Weak	97	93
M	E	P	Strong	2	0
M	M	C		1	3
M	M	E		185	183
M	M	P		71	75
Total strongly intransitive ratings				10 (0.9%)	14 (1.3%)
Total weakly intransitive ratings				224 (21.4%)	186 (17.2%)
Total intransitive ratings				234 (22.3%)	200 (18.5%)

The letter C, M or P indicates a preference for that procedure (CABG, MED, PTCA, respectively), while E indicates equivocation in the comparison.

Each panellist made between 1000 and 1100 three-part ratings per round. On round 1, panellists 2 and 6 had very few strongly intransitive ratings, while panellists 2, 11 and especially 3 were high (5–10%). Most panellists had between 20 and 32 strongly intransitive ratings (> 3%). In round 2, six panellists increased their frequency of strongly intransitive and six decreased; in this round, six panellists had more than 3% strongly intransitive. With respect to weak intransitivity, the panellists split on round 1 into a low group of seven members (range 11–114 or 1–10%) and a high group of five members (range 183–270 or 16–25%). No member switched groups between round 1 and 2, although three members formed a middle group around 13% weakly intransitive. Propensity to form strongly intransitive combinations did not appear to be related to the propensity to form weakly intransitive patterns.

5. Conclusion and discussion

In general, the results of the two rounds showed reasonable consistency, but also considerable and important changes between rounds.

First, there was a substantial increase in the degree of agreement amongst panel members together with a corresponding decrease in disagreement. Such an increase was to be expected as members could listen to the opinions of others and their understanding of the panel process grew. Similar changes have been noted by Park et al. [2] and Leape et al. [9]. Park et al. [2] rated six medical and surgical procedures, using four different definitions of agreement. Substantial increases in agreement were observed for four procedures, with little change for the other two.

Second, there was a decrease in the number of indications rated equivocal by the panel with a movement towards both appropriate and inappropriate. This change was mainly due to changes in the median opinion of the panel and was only slightly due to the decrease in disagreement. Again, this parallels some but not all earlier findings. The fact that approximately 90% of all indications were rated identically over the two rounds shows that the appropriateness ratings are very stable.

Third, most individual members scored relatively few indications strongly intransitively, indicating that they generally had a good understanding of the concept of the comparisons. The panel as a group scored very few strongly intransitive combinations and these were largely concentrated in indications involving a C lesion, reflecting the perceived difficulties involved in treating a type C lesion.

Finally, several members appear to have undergone a process of learning during the first round. They changed many of their ratings quite dramatically, possibly due to misunderstandings of procedures or misunderstandings in the first round.

Although changes were observed between the two rounds, the overall stability of ratings indicates that a third round would not lead to further substantial changes. Changes due to misunderstandings would disappear and there is no indication that intransitivities would decrease.

In summary, the observed changes in appropriateness ratings were consistent with expectations both in magnitude and direction. The ratings obtained show stability and are consistent with perceived medical thought. The observed panellist changes show the advisability of multiple rounds of ratings. Based on an internal analysis of the DUCAT study, the modified Delphi process offers a consistent method of determining the appropriateness of alternative treatments.

Table 10
Number of intransitively scored indications per panel member per round

Member	Round	1	2	3	4	5	6	7	8	9	10	11	12
Strong	1	32	54	100	31	23	7	21	31	30	27	69	2
	2	39	77	42	22	18	8	19	19	34	42	33	4
Weak	1	183	183	50	192	11	82	201	201	270	89	94	114
	2	147	147	71	141	41	89	248	248	192	127	111	142

Appendix A: Marginal symmetry

The P-values reported in Section 3 are those of a test of marginal symmetry in a square contingency table. As this test may not be familiar to many practitioners, we include a short description of it. Further details may be found in Agresti [10].

The results of the comparison of the two rounds were displayed in a 3 × 3 contingency table e.g. the appropriateness rating for PTCA (vs. medical therapy) was given in Table 4. More generally, such a table can be represented as

n_{11}	n_{12}	n_{13}	n_{1+}
n_{21}	n_{22}	n_{23}	n_{2+}
n_{31}	n_{32}	n_{33}	n_{3+}

where $n_{i+} = n_{i1} + n_{i2} + n_{i3}$ for $i = 1$–3, and n_{+j} is similarly defined ($j = 1$–3).

The null hypothesis is that the distribution of the appropriateness ratings is the same for rounds 1 and 2 i.e. $n_{i+} = n_{+i}$ for $i = 1$–3 against the alternative hypothesis that $n_{i+} \neq n_{+i}$ for some i.

Note that since $n_{1+} + n_{2+} + n_{3+} = n_{+1} + n_{+2} + n_{+3} = n_{++}$, one of the equations involved is redundant.

Clearly, if the null hypothesis is true, the vector **d**, where **d** has elements $d_i = (n_{i+} - n_{+i})/n_{++}$ has expected value *0*. The vector **d** has length 2 because of the redundancy noted above. Under the null hypothesis, the estimated variance matrix of **d** has elements $w_{ii} = (n_{i+} + n_{+i} - 2n_{ii})/n_{ii}$ ($i = 1,2$) and $w_{ij} = -(n_{ij} + n_{ji})/n_{++}$ ($i = 1,2; j = 1,2; i \neq j$).

The test statistic is then $Q = n_{++}\,\mathbf{d}'\mathbf{W}^{-1}\mathbf{d}$ which under the null hypothesis has a chi-squared distribution with two degrees of freedom.

References

[1] Kahn, K.L., Kosecoff, J., Chassin, M.R. et al., Measuring the clinical appropriateness of the use of a procedure, Medical Care, 26 (1988) 415–422.

[2] Park, R.E., Fink, A., Brook, R.H. et al., Physician ratings of appropriate indications for six medical and surgical procedures, American Journal of Public Health, 76 (1986) 766–772.

[3] Hilborne, L.H., Leape, L.L., Bernstein, S.J. et al., The appropriateness of use of percutaneous transluminal coronary angioplasty in New York State, Journal of the American Medical Association, 269 (1993) 761–765.

[4] Rigter, H., Appropriate use of the concept 'appropriate' in discussions on health care (in Dutch), Nederlands Tijdschrift voor Geneeskunde 138(1) (1994) 4–7.

[5] Meijler, A.P., McDonnell, J. and Rigter, H., Assessment of indications using the RAND method: invasive therapy in coronary sclerosis as an example (in Dutch), Nederlands Tijdschrift voor Geneeskunde, 138(1) (1994) 22–28.

[6] Hilborne, L.H., Leape, L.L., Kahan, J.P., et al., Percutaneous Transluminal Coronary Angioplasty: A Literature Review and Ratings of Appropriateness and Necessity, Report JRA-01, RAND, Santa Monica, CA, 1992.

[7] Leape, L.L., Hilborne, L.H., Kahan, J.P., et al., Coronary Artery Bypass Graft: A Literature Review and Ratings of Appropriateness and Necessity, Report JRA-02, RAND, Santa Monica, CA, 1992.
[8] Winslow, C.M., Kosecoff, J.B., Chassin, M.R., et al., The appropriateness of performing coronary artery bypass surgery, Journal of the American Medical Association, 260 (1988) 505–509.
[9] Leape, L.L., Freshour, M.A., Yntema, D., et al., Small-group judgement methods for determining resource-based relative values, Medical Care, 30 (1992) NS28–39.
[10] Agresti, A., Categorical Data Analysis, Wiley, New York, 1990, pp. 365–370.