Neurourology and Urodynamics 13:323–331 (1994)

Statistical Evaluation of the Function of the 1992 International Continence Society Scientific Committee

R. van Mastrigt and J.W. Downie

Department of Urology-Urodynamics, Erasmus University Rotterdam, Rotterdam, the Netherlands (R.v.M.); Department of Pharmacology, Dalhousie University, Nova Scotia, Canada (J.W.D.)

Papers submitted to the International Continence Society are read by the six members of the scientific committee, who assign three scores to each paper, one for its originality, one for its scientific value, and one for its academic or clinical interest. Following discussion at the scientific committee meeting, a program is made from the abstracts with the highest scores. In this study statistical properties of the assigned score values for the 1992 ICS meeting are discussed. It is concluded that the three scores do not measure independent properties of the abstracts, that different abstract types (for instance "clinical" and "basic" papers) are scored differently by the different committee members, that there is a significant consensus among the committee members, that the scientific committee meeting has a relatively small effect on the scores, and that different abstract types are not equally represented in the final program. © 1994 Wiley-Liss, Inc.

Key words: International Continence Society, abstracts, abstract scores, scores, scientific meeting

INTRODUCTION

Abstracts submitted for ICS meetings are collected by that specific years' scientific chairman and anonymous versions of it are sent in the order in which they are received to the members of the scientific committee. Apart from the chairman this committee consists of the chairman of the previous year, the chairman of the next year, one local representative, and two ICS representatives, one scientific and one clinical representative that are alternatingly elected by the ICS for a 2 year period. All six committee members read all abstracts and assign three scores to each abstract [except to those they (co)author]: one for its originality, one for its scientific value, and one for its academic or clinical interest. Each score ranks from 0 to 3. The chairman collects all scores in a database. After the deadline for abstract submittal has passed the committee meets and discusses (still anonymous) all abstracts, with special attention to those abstracts that show wide discrepancies in scoring. At this stage

Received for publication October 4, 1993; accepted January 18, 1994.

Address reprint requests to Dr. ir. R. van Mastrigt, Department of Urology-Urodynamics, Room EE1630, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, the Netherlands.

committee members can change the assigned scores. All scores are added and finally a program is made up from the abstracts with the highest total scores. The cut-off point for presentations is not predefined but derived from the distribution of scores and the number of available slots in the program.

At the 1992 scientific committee meeting some questions emerged regarding the scoring system that might be answered by statistical processing of the assigned scores. In this article the following aspects will be discussed:

1. Are the scores normally distributed, and/or should they be normally distributed?
2. Are the three scores (for originality, scientific value, and clinical or academic interest) mutually independent?
3. Are the scores dependent on the category of the abstract and the background of the committee members? (for instance, do "basic science" papers score differently from "clinical papers" and are there differences between the scores from basic scientists and clinicians?)
4. Does the scoring change as a function of time, i.e., are abstracts submitted, and therefore read, later scored differently from earlier submitted abstracts?
5. How are the scores affected by the discussion at the scientific committee meeting; does this increase the degree of consensus?
6. How are the abstract categories distributed over the program categories (for instance, are "clinical" papers rejected more often than "basic" papers)?

METHODS

Following the ICS meeting both the scores assigned before and after the scientific committee meeting were transferred to an IBM compatible PC and were processed using the statistical package SPSS. Apart from the three scores from each committee member and the abstract number (which is a number given when the abstract was received, and therefore represents the historical order in which abstracts are read by the committee members), an abstract type number was assigned. Abstracts involving isolated tissue or using experimental animals were called "basic." Urodynamic studies were called "clinical" unless the thrust was mathematical or the study was intended to compare methodologies or explore new methodologies or technologies in which case they were considered "basic urodynamics." All other abstracts involving patients or patient-related issues were labelled "clinical" except for survey-based studies which were considered a separate category ("survey"). Numbers between 1 and 6 were assigned randomly to anonymously represent the committee members. If not explicitely stated otherwise, all results shown refer to the score values after the scientific committee meeting, i.e., after the committee members had had an opportunity to change their scores as a result of the discussions in the meeting.

RESULTS

In 1992 324 abstracts were received and scored. Table I shows for each of the six members of the scientific committee the distribution of the scores for each of the three categories: originality, scientific value, and interest. Also the mean, standard deviation, and skewness of the distributions are shown, as well as the significance of

TABLE I. The Distribution of Scores for Each of the Six Members of the Scientific Committee

Member	Category[a]	Frequency of scores 0	1	2	3	Mean	S.D.	Skew	Significance
1	Ori	13	134	139	33	1.6	.73	.13	.000
	Sci	10	107	183	20	1.7	.64	−.28	.000
	Int	2	156	141	20	1.6	.62	.48	.000
2	Ori	9	85	182	47	1.8	.70	−.24	.000
	Sci	4	87	180	52	1.9	.68	.07	.000
	Int	4	110	184	25	1.7	.62	.03	.000
3	Ori	43	150	119	6	1.3	.71	−.15	.000
	Sci	30	155	130	3	1.3	.66	.27	.000
	Int	20	191	103	4	1.3	.60	.16	.000
4	Ori	31	137	127	27	1.5	.78	.04	.000
	Sci	39	140	131	12	1.4	.74	−.14	.000
	Int	29	145	130	18	1.4	.73	.00	.000
5	Ori	71	105	115	24	1.3	.90	.02	.000
	Sci	112	109	83	11	1.0	.87	.36	.000
	Int	75	128	106	6	1.1	.80	−.02	.000
6	Ori	9	77	187	43	1.8	.68	−.32	.000
	Sci	4	132	170	10	1.6	.58	.05	.000
	Int	3	137	161	15	1.6	.60	.17	.000

[a]Ori, originality; Sci, scientific value; Int, interest.

the difference from a normal distribution according to the Kolmogorov-Smirnov Goodness of Fit Test. It can be seen that all scores from all members were not normally distributed. The test compares the tested distribution with a normal distribution with the same mean and standard deviation. The fact that for all scores, except the originality score from member 4, the mean was not 1.5 (the midpoint of the score range) therefore did not contribute to the significance of the test. There is no uniform pattern in the abnormality of the score distribution. Some distributions are strongly skewed to the right; see for instance the originality score of member 6; in addition to the fact that the mean of this distribution is larger than 1.5, there are many more papers that received a score above the mean than papers that had a score below the mean. Other distributions are strongly skewed to the left; for instance, the scientific value score of member 5; in spite of the low mean value of 1.0 many more papers scored below the mean than above.

Table II shows the degree to which the three scores of each committee member were correlated. In fact, all scores were significantly correlated for all members. The smallest correlation was found between the originality and interest score of member 6 (0.10) and the highest between originality and interest of member 5 (0.63).

Table III gives correlations among the scores of different committee members. With a few exceptions the scores were significantly correlated. Without exception the interest scores were least correlated, i.e., the least agreement among committee members existed on the interest score. As stated in "methods" the values in the table reflect the score values following the scientific committee meeting. A similar table based on the values before the meeting looked almost identical; none of the displayed values differed more than ± 0.01 between the two tables.

TABLE II. The Correlations Between the Three Scores for Each of the Six Members of the Scientific Committee, and the Associated Significances (Pearson Correlation)

Member	Originality/ Scientific value	Signifi- cance	Scientific value	Signifi- cance	Originality/ Interest	Signifi- cance
1	.48	.000	.59	.000	.51	.000
2	.29	.000	.29	.000	.24	.000
3	.38	.000	.26	.000	.23	.000
4	.32	.000	.47	.000	.42	.000
5	.51	.000	.55	.000	.63	.000
6	.36	.000	.18	.001	.10	.044

TABLE III. Pearson Correlations Between the Scores of the Different Committee Members*

Member	Score[a]	2	3	4	5	6
1	Ori	.36	.43	.37	.35	.33
	Sci	.28	.41	.25	.49	.37
	Int	.10(−)	.13(−)	.27	.31	.11(−)
2	Ori		.37	.27	.36	.32
	Sci		.34	.32	.44	.31
	Int		.22	.26	.05(−)	.22
3	Ori			.34	.35	.39
	Sci			.34	.55	.32
	Int			.14	.06(−)	.22
4	Ori				.39	.33
	Sci				.43	.27
	Int				.22	.16
5	Ori					.35
	Sci					.38
	Int					−.02(−)

*All correlations are significant at the 1% level except those indicated with (−).
[a]Ori, originality; Sci, scientific value; Int, interest.

Table IV gives a breakdown of mean score values for the different abstract types. The significance of Pearson's chi square indicates whether scores were equally distributed over the abstract types. Only committee member 4 had scores that were independent of the abstract types; for all other members there was a significant correlation between the abstract scores and the type of abstract. In the two sets of rows marked "Basic scientists" and "Clinicians," the average scores of the two, respectively four, committee members with this background were added. Except for the interest scores of the basic scientists, all the scores were significantly dependent on the abstract type. Moreover, with two exceptions (the originality score for "basic" abstracts and the interest score for "basic urodynamics" abstracts) the score values of basic scientists and clinicians were significantly different (Wilcoxon matched pairs test $P = 0.05$).

Table V shows the rank correlations between the scores and the abstract numbers. As abstract numbers were assigned in historical order, this correlation is a measure of the change in time of the scoring of the committee members. Only two correlations were significant; both were positive and for the originality score, implying that both these committee members tended to score later abstracts higher for

TABLE IV. The Mean Scores of All Members for the Different Abstract Types, and the Significance of Pearson's Chi Square

		Abstract type				
Member	Category[a]	Basic (N = 57)	Basic urodynamics (N = 17)	Clinical (N = 231)	Survey (N = 19)	Significance
1	Ori	1.7	2.3	1.5	1.6	.00000
	Sci	2.0	2.2	1.5	1.7	.00000
	Int	1.5	2.1	1.5	1.7	.0002
2	Ori	2.1	2.3	1.8	1.4	.00000
	Sci	2.2	2.4	1.8	1.6	.00002
	Int	1.5	1.5	1.8	1.6	.05
3	Ori	1.7	1.7	1.1	1.4	.00003
	Sci	1.9	1.4	1.2	1.4	.00000
	Int	1.2	1.2	1.3	1.7	.006
4	Ori	1.6	1.8	1.4	1.4	.12
	Sci	1.5	1.4	1.3	1.4	.25
	Int	1.3	1.4	1.4	1.6	.36
5	Ori	2.0	1.4	1.2	.7	.00000
	Sci	1.8	1.1	.8	.8	.00000
	Int	1.7	1.1	1.0	.6	.00000
6	Ori	2.1	1.9	1.8	1.9	.003
	Sci	1.9	1.9	1.5	1.6	.00007
	Int	1.2	1.3	1.7	1.9	.00000
Basic scientists[b]	Ori	1.8	1.5	1.3	1.1	.0001
	Sci	1.6	1.2	1.0	1.1	.00000
	Int	1.5	1.2	2.2	1.1	.069
Clinicians[b]	Ori	1.9	2.1	1.5	1.5	.0001
	Sci	2.0	2.0	1.5	1.5	.00000
	Int	1.4	1.5	1.6	1.7	.00001

[a]Ori, originality; Sci, scientific value; Int, interest.
[b]Data in Basic Scientists and Clinicians show averages of the two and four committee members, respectively, with this background.

originality. Table VI gives the distribution of score changes following the committee meeting; 3.2% of the scores was changed in the meeting. The changes affected 22% of the abstracts as shown in Table VII. Finally Table VIII shows the distribution of the abstract categories over the program categories. Different abstract categories were significantly differently represented in the program categories according to Pearson's chi square.

DISCUSSION

The data shows that none of the scores of any member of the scientific committee was normally distributed. It can be wondered if a normal distribution was to be expected. Such a distribution arises from random variation of a variable around a mean. In the analysed scores there are two sources of variation. One is the intrinsic variation, i.e., the abstract property that the scores attempt to quantify is different for each abstract. The other source of variation is the committee members' estimation of

TABLE V. The Rank Correlations of the Scores of All Members With the Abstract Number

Member	Category[a]	Rank correlation	Significance
1	Ori	.1838	*
	Sci	−.0083	
	Int	−.0165	
2	Ori	−.0387	
	Sci	.0141	
	Int	−.1343	
3	Ori	.1155	
	Sci	.0691	
	Int	−.0078	
4	Ori	.2260	*
	Sci	−.0471	
	Int	.0978	
5	Ori	−.0790	
	Sci	−.0279	
	Int	−.0597	
6	Ori	.0736	
	Sci	.0085	
	Int	−.0470	

[a]Ori, originality; Sci, scientific value; Int, interest.
*Significant at the 1% level.

the property. It is not unlikely that this last variation can be described as a random fluctuation of the scores, but it is not likely that the intrinsic variation is random. This would imply for instance that the likelihood that authors would submit a very original abstract is the same as the likelihood that they would submit an unoriginal abstract. Probably the latter is much easier and therefore will happen more frequently. The data in Table I reflects the combined effect of the two sources of variation, intrinsic and committee-member variation. As there are large differences between the resulting distributions—approximately one half is skewed to the left and the other half is skewed to the right—this must be ascribed to committee member variation. It is probably not possible to draw conclusions on the intrinsic variation, or the intrinsic distribution of the scores. In spite of the considerable committee member variation there is abundant common ground, which can be seen in Table III. Most of the scores of most of the members are significantly positively correlated. These significant correlations between the members scores are not caused by the discussion at the scientific meeting. This discussion resulted in a relatively small number of changes, in 3.2% of the scores only, affecting however a considerable number of abstracts, i.e., 22%. The changes were equally distributed over the abstract categories (Table VII), i.e., committee members changed their score values for "clinical" papers as often as for "basic" papers etc. relating to the relative number of papers in each category. About half of the changes in scores were changes in the originality score; in seven abstracts (0.12%) this score was changed by 3 points, i.e., from maximally original to maximally unoriginal. The originality score is obviously the most dependent on specific knowledge or background and therefore the most sensitive to discussion. The least

TABLE VI. The Distribution of the Changes in Scores Following the Scientific Meeting

Member	Category[a]	Number of changes in scores with magnitude −3	−2	−1	+1	+2	Total
1	Ori		4	8			12
	Sci			6			6
	Int			2	1		3
2	Ori	2	4	9	2	1	18
	Sci		2	7	1	2	12
	Int		2	3	4		9
3	Ori	1	1	4	3		9
	Sci			1	3		4
	Int			1	4		5
4	Ori		5	11	5	1	22
	Sci			9		1	10
	Int		1	9	1		11
5	Ori		3	9	5	3	20
	Sci			3	1	1	5
	Int			2	4	2	8
6	Ori	4	3	9	1		17
	Sci			5	1		6
	Int		1	6	1		8
Total		7	26	104	37	11	185
Percentage[b]		0.12	0.45	1.7	0.63	0.19	3.2

[a]Ori, originality; Sci, scientific value; Int, interest.
[b]Percentage gives the number of times a change of certain magnitude occurred as a percentage of the total number of scores, which was: 324 (abstracts) × 3 (scores) × 6 (members) = 5,832.

TABLE VII. The Number of Abstracts in the Four Abstract Categories for Which One or More of the Scores Was Changed at the Scientific Committee Meeting*

	Basic	Basic urodynamics	Clinical	Survey	Total
No change	43	15	179	16	253
Change	14	2	52	3	71
Total	57	17	231	19	324

*Pearson's chi square = 1.72, $P = 0.63$.

agreement existed on the interest score. Table III shows that in 6 out of the 15 possible member combinations there was no significant agreement on this score. With one marginal exception the correlation of the interest score was the lowest in all member combinations. This probably reflects that the interest score is the most subjective of the three scores. For each member there was a significant correlation between the three scores with, depending on the significance level, one possible exception (the originality and interest score of member 6; see Table II). These high correlations

TABLE VIII. The Number and Percentages (in parentheses) of Abstracts in Each of the Four Abstract Categories in the Seven Program Categories*

	Podium	Formal poster	Informal poster	Video	Withdrawn	Read by title	Reject
Basic	3 (5%)	32 (56%)	14 (25%)			6 (10%)	2 (4%)
Basic urodynamics	7 (41%)	2 (12%)	3 (18%)	1 (6%)		4 (24%)	
Clinical	24 (10%)	40 (17%)	42 (18%)	8 (4%)	1	116 (50%)	
Surveys	2 (10%)	6 (32%)	1 (5%)			10 (53%)	
Total	36 (11%)	80 (25%)	60 (18%)	9 (3%)	1	136 (42%)	2 (1%)

*Pearson's chi square = 80, $P = 0.000005$.

signify that the three scores are not independent, they do not quantify clearly different properties of the abstracts, and none of the committee members was able to independently and differently score these properties. It follows that if it is thought desirable to attempt to independently measure different aspects of the abstracts, it is necessary to use other scores that quantify more clearly defined, separate aspects of the abstracts. The alternative is to use only one overall score for each abstract.

Five of the six members of the scientific committee scored different types of abstracts differently. For these five members, not only the interest score but also the appreciation of originality and scientific value was clearly different for the four abstract categories. The fact that the three scores are not truly independent also may play a role here. When the committee members were grouped as "basic scientists" vs. "clinicians" 10 of the 12 score values in the four abstract categories were significantly different between these groups. This difference in scoring justifies the composition of a committee composed of members with different background. It should be noted that one committee member systematically managed to avoid this background based bias.

In the final program not all abstract categories were equally represented. "Clinical" papers and "surveys" were more often excluded from presentation (read by title) than "basic" and "basic urodynamics" papers. On the other hand three times more of these papers were submitted than "basic" and "basic urodynamics" papers, and two times more clinical papers were accepted to the program then basic papers. Compared to "clinical" papers "basic" papers were less frequently presented on the podium, more frequently as a poster, and less frequently as read by title. "Basic urodynamics" papers were more often presented on the podium than "clinical" papers and less frequently as "read by title." This distribution of the abstracts over the program is only partly determined by the number of accepted abstracts within topic groupings. The decisions about podium versus formal poster presentation are a major component of the scientific committee's deliberations at its meeting. Although these decisions are made with the committee still "blinded" to the identity of the authors, they are influenced by the members' perceptions of what sort of material is best presented orally or by static display.

A final surprising finding was that with two exceptions the scorings did not significantly change over time and that the two exceptions were positive and for the same score. Two committee members tended to appreciate abstracts as more and more original while reading 324 of them.

In conclusion it can be stated that in the data from the 1992 Scientific Committee of the ICS:

1. The abstract scores were not distributed normally.
2. The three scores that were supposed to measure independent properties of the abstracts were not independent. This situation may be remedied by changing the score definitions, scoring other aspects of the abstracts, or by using a single overall score.
3. Different types of abstracts (e.g., "basic" vs. "clinical") were generally scored differently by different members or by the basic scientists versus the clinicians. In 1992 one member managed to avoid this bias.
4. With two (positive!) exceptions scores did not change over time, as the committee members read more and more abstracts.
5. 3.2% of the score values were changed at the scientific committee meeting, affecting 22% of the abstracts, irrespective of the abstract category. The degree of consensus between the committee members did not change significantly as a result of the discussion.
6. Different abstract categories were distributed differently over the program categories. Compared to "clinical" papers, "basic" papers were less frequently presented on the podium, more frequently as a poster, and less frequently as read by title. "Basic urodynamics" papers were more often presented on the podium than "clinical" papers and less frequently as "read by title."

ACKNOWLEDGMENTS

The authors are grateful to the other committee members, W. Artibani, L.D. Cardozo, J.B. Gajewski, and K. Höfner, for their kind permission to use the score data for this paper.