MEDICAL PRACTICE

Contemporary Themes

Observer variation in assessment of results of surgery for peptic ulceration

ROBERT HALL, JANE C HORROCKS, SUSAN E CLAMP, F T DE DOMBAL

British Medical Journal, 1976, **1**, 814-816

The results of surgery for peptic ulcer may be assessed in many ways. For instance, an assessment may focus on such factors as mortality and morbidity immediately after operation. Early postoperative mortality is, however, low—whatever the operation—and most workers have therefore concentrated on the medium and long-term results of surgery and have dealt with the presence or absence of symptoms attributable to recurrent ulceration or the operative procedure itself, together with the severity of these symptoms. The patient's overall status has usually also been graded, the most popular grading system being that devised by Visick.[1] As little has been done to assess the reproducibility of these methods of assessment we carried out observer variation studies in a series of 170 patients seen at gastric follow-up clinics in York. We report here our findings.

Patients and methods

170 patients attending clinics in the County Hospital, York, after surgery for peptic ulcer were studied. All patients had undergone surgery for peptic ulceration between 1944 and 1974 and were attending for routine yearly review.

In the follow-up clinic each patient was reviewed (as was normal practice) by a panel of clinicians who were unaware of the operation performed. This usual review was carried out in each case by two senior clinicians (RH and Mr D Johnston), and a similar review was carried out by a second panel comprising the three remaining authors of this report. The reviews of each panel were compared.

Each patient was assessed in terms of (*a*) the presence or absence of the following symptoms (together with their severity where relevant): pain, nausea, vomiting (bile or food), appetite, epigastric fullness, diarrhoea, dumping (early or late), reflux, heartburn, flatulence, dysphagia; and (*b*) the overall Visick grading.

Visick classification—The original classification put forward by Arthur Hedley Visick[1] is as follows: grade I, no symptoms; grade II, mild symptoms relieved by care; grade IIIs, symptoms not relieved by care, but the patient's overall result is satisfactory; IIIu, symptoms not relieved by care and the result is unsatisfactory; grade IV, no improvement. Grades IIIu and IV are considered to represent failure. Most surgeons have now adopted a modification of Visick's original classification and that currently used in the gastric follow-up clinic in York is shown below.[2]

Modified Visick classification—Grade I, absolutely no symptoms. perfect result; grade II, patient considers results perfect, but interrogation elicits mild occasional symptoms easily controlled by minor adjustments to diet; grade III, mild or moderate symptoms not controlled by care, causing some discomfort, but patient and surgeon satisfied with result, which does not interfere seriously with life or work; grade IV, moderate or severe symptoms or complication that interferes considerably with work or enjoyment of life; patient or doctor dissatisfied with result. Includes all cases with proved recurrent ulcer and those submitted to further operation, even though second operation may result in considerable symptomatic improvement. In this modified system there are changes in detail but not in concept, the most notable change being a reduction to four categories, of which grades I and II are considered to be good or excellent, grade III moderate or fair, and grade IV to represent outright failure of surgery. In our study we initially used the modified form of Visick's classification.

Results

At the beginning of the study both panels agreed on the presence or absence of symptoms in 93·7% of cases and on the severity of

York Peptic Ulcer Research Group, County Hospital, York

ROBERT HALL, FRCS, consultant surgeon

JANE C HORROCKS, computer programmer and physicians' assistant, University Department of Surgery, Leeds

SUSAN E CLAMP, assistant in surgical research, University Department of Surgery, Leeds

F T DE DOMBAL, MD, FRCS, reader in clinical information science, Leeds University.

TABLE I—*Initial findings in gastric follow-up clinic: consensus results between two panels*

Aspect of assessments	No of assessments	Agreed	Disagreed	% Agreement
Symptoms present or absent	221	207	14	93·7
Severity of symptoms	221	202	19	91·4
Visick gradings	17	11	6	64·7

individual symptoms (when applicable) in 91·4% (table I). Agreement on the patients' overall grading on the Visick scale, however, was much less close—only 64·7%—which indicated that in roughly a third of the patients studied the consensus views of the two panels differed in the overall Visick gradings.

As a result of these preliminary findings, the members of both panels discussed the precise nature of the Visick classification used. The effect of this discussion is shown in table II. In the first two gastric-follow-up clinics the variation on the presence or absence of symptoms was 5·1%, on the severity of symptoms 8·7%, and on the overall Visick grading 26·7%. During the next two clinics the figures remained roughly comparable, but thereafter a decided decrease was noticed in the variation in all three categories of observation.

TABLE II—*Summary of data on observer variation in series of 170 cases*

	No of patients seen	Degree of observer variation on:		
		Presence or absence of symptoms	Severity of symptoms	Visick gradings
1st Clinic	17	6·3%	8·6%	35·3%
1st and 2nd Clinics	30	5·1%	8·7%	26·7%
3rd and 4th Clinics	33	4·0%	6·3%	24·2%
5th-10th Clinics	107	1·7%	3·4%	9·3%

Over the next six clinics (107 patients) the variation between observers on the presence or absence of symptoms was only 1·7%, on severity only 3·4%, and on Visick grades 9·3%. Thus although initial Visick gradings differed a high degree of agreement was reached (albeit only after three to four months of discussion).

Original v modified Visick gradings—One possible reason for the problem of assessing patients under the Visick scale might have lain in the fact that the original Visick classification was not used in York. For an additional 40 cases, therefore, the same experiment was repeated, but panel members were asked to assess patients on both the original Visick system and the more commonly used modified system. There was virtually no difference between these alternative modes of assessment overall agreement being 82·5% for the original classification and 85% for the modified version.

Clinical v non-clinical assessments—Another obvious potential explanation for these findings lay in the disparity of experience among the members of the panels, particularly because only one member of the second panel was medically qualified. For the first two clinics, however, the findings were strictly comparable between clinical (65% agreement) and non-clinical (71% agreement) members of the panels. Moreover, overall levels of agreement between differing combinations of individual participants were similar throughout the trial, so disparity of experience among participants did not seem to be responsible for the findings in table I.

Analysis of Visick gradings—Possibly one particular problem in classification was responsible for the differences in Visick gradings (such as the difference between grades II and III), but, as shown in table III, the variation in Visick gradings was more or less equally distributed among the four categories used.

TABLE III—*Comparison of Visick Gradings by panels 1 and 2 in 170 cases*

Visick grading by panel 2	Visick grading by panel 1				
	I	II	III	IV	Total
I	43	5			48
II	6	54	4		64
III	1	4	39	2	46
IV			4	8	12
Total	50	63	47	10	170

Analysis of individual symptoms—Again, problems with one or two individual symptoms might have been responsible for the variation in the analysis of symptoms. But no individual symptom was recorded without some disagreement, and although observer variation on individual symptoms ranged from 0·7% to 11·2% the frequency of variation was related to the frequency of the symptom itself. Thus the "worst" symptoms in respect of observer variation (epigastric fullness 11·2%; nausea 10·0%) were also the most common in these (possibly unrepresentative) patients. Thus no single symptom or Visick grading was responsible for the variation among observers.

FURTHER STUDIES

We decided to introduce a further variable into the assessment—namely, the patients' own assessment of the outcome of surgery. Each patient was asked (before interview) whether he considered the outcome of treatment a tremendous success, a decided improvement, or a failure, or whether he was not sure whether the operation was worth having or not. The patient was also asked to rate his own status by placing a mark on a line (a visual linear analogue scale) stretching from "awful" to "perfect."[3] The results of these two methods of assessment were then compared with the patient's Visick grading.

The patients' own two responses agreed closely (fig 1); most patients who indicated that their operation had been a tremendous success placed their mark on the linear analogue scale towards the perfect end of the line, and there was a reasonably good correlation with the other forms of assessment. The correlation between the patients' response on the linear analogue scale and the Visick gradings given to them in the clinic was, however, much poorer (fig 2). In particular, most patients who were graded Visick III considered themselves to be either perfect or very nearly so. This was confirmed by the fact that while all 35 patients graded Visick I considered their operation to be a tremendous success, so did well over half the patients graded Visick II or III. Indeed all patients graded Visick III considered themselves decidedly improved as a result of surgery.

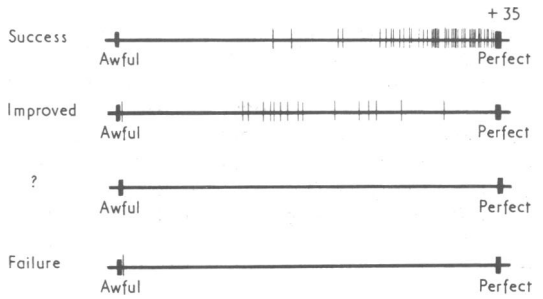

FIG 1—Correlation between two forms of patient assessment (patients' own overall grading and patients' mark on linear analogue scale). Figure at right indicates numbers of patients placing marks at "perfect" end of scale.

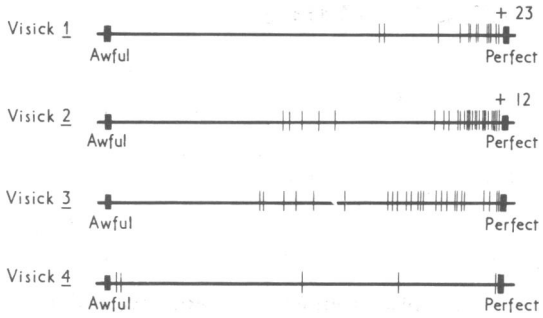

FIG 2—Correlation between overall Visick grading and patients' own rating on the linear analogue scale. Figure at right indicates numbers of patients placing marks at "perfect" end of scale.

Discussion

Clearly the most reliable method of assessing a patient's clinical status after surgery for peptic ulcer is to assess the presence or absence of specific symptoms. The overall observer

agreement in respect of this was excellent and these overall findings held good for each observer. It has been claimed that it is difficult to assess the presence and severity of, for example, diarrhoea,[4] but we found that the variation between observers was only 1·8% for the presence of diarrhoea and only 6% for its severity. Comparisons between operations, or between centres, in terms of the presence or absence of symptoms are therefore not likely to be invalidated by large variations between observers. Unfortunately, the same cannot be said of the overall assessment by Visick's system of grading. The initial variation among observers in recording this grade (both between qualified and non-qualified observers) was unacceptable. It could be argued, of course, that this merely reflected the inexperience of some members of the assessing panels, but the initial observer variation was not person specific, and each member of both panels had spent at least one year carrying out research in gastric follow-up clinics before the study; indeed each of the clinical members had for several years taken part in follow-up studies after gastric surgery. Hence, whatever its intrinsic merits, Visick's grading system is inappropriate for comparing results from different surveys or centres, unless the observers in both centres have undertaken joint observer variation trials and discussed the system between themselves.

The Visick classification is not, however, valueless. Certainly when it first appeared in 1948 it was an immense improvement on anything that had existed before then, and it is a perfectly valid means of assessment by which workers in a specific centre may compare the results of different operations. Nevertheless, even used in this way, Visick's grading is rather imprecise in that small differences between operations may pass unnoticed owing to the intrinsic observer variation in the method of assessment; this is like the adverse signal:noise ratio that is familiar to electronic engineers: the "signal" (the difference between operations) is lost in the "noise"(the intrinsic variation between observers). One would therefore expect, in the light of our findings, to find that the results of any trial would show the incidences of Visick grades for various operations all lying within a few per cent of each other. Perhaps it is not too far fetched to assert that this has been the experience of British trials of surgery for peptic ulcer over the past 30 years.

There is some evidence (from figs 1 and 2) that the Visick grades do not even correspond to patients' own feelings about the results of their surgery. This was particularly so in patients graded Visick III (because they cannot control their symptoms), who regarded themselves as either perfectly fit or very nearly so. This finding rather bears out Visick's concept that patients could be graded Visick III at a particular visit yet overall may be regarded as "satisfactory" and rather mitigates against the more modern trend of expressing the "success rate" of an operation in terms of the proportion of patients graded Visick I or II. Which is the "right" assessment is, for the moment, an open question, but the discrepancy exists and is worthy of further study.

We make three specific suggestions. Firstly, the use of Visick grading should be limited to comparisons made in single centres between different operations using single teams of observers. Secondly, comparisons between different centres (or between different teams of observers in the same centre) should concentrate on the presence or absence of specific symptoms; whatever the deficiencies of this form of analysis, it does at least seem to be more reliable and reproducible. Thirdly, this is an appropriate time to examine alternative means of assessing patients. We have described one such method in this report—using linear analogue scales. We have also carried out studies with "state flow charts," which are to be reported elsewhere.[5] Psychiatric assessments of patients' responses have been analysed,[6] and, clearly, the patients' own apparent responses are often at variance with the response elicited during the "traditional" clinical follow-up interview. Any future modification of this traditional "symptomatic" assessment should, however, remain simple, practical, and, above all, reproducible.

We thank Mr D Johnston for his participation in this trial as an observer and for helpful comments and advice during the preparation of this paper. We thank Mrs Mary Dent and Mrs Sheila Dickson for their help in organising the follow-up clinics. JCH was aided by a grant from the Medical Research Council, which we acknowledge with gratitude. Finally, and particularly, we thank the patients who participated and gave us their time and their helpful comments in the conduct of this study.

References

1 Visick, A H, *Annals of the Royal College of Surgeons of England*, 1948, **3**, 266.
2 Goligher, J C, *et al*, *British Medical Journal*, 1968, **2**, 781.
3 Huskisson, E C, *Lancet*, 1974, **2**, 1127.
4 Cox, A G, and Bond, M G, *British Medical Journal*, 1964, **1**, 460.
5 Horrocks, J C, *et al*, in preparation.
6 Cay, E L, *et al*, *Lancet*, 1975, **1**, 29.

Causes of failure to harvest cadaver kidneys for transplantation

A McL JENKINS

British Medical Journal, 1976, **1**, 816-817

Summary

Fifty-two possible donors of cadaver kidneys were referred to the Nuffield Transplantation Surgery Unit, Edinburgh, in 12 months. Only 12 (23%) yielded kidneys, while a further 12 were medically unsuitable as donors. Refusal by relatives to allow cadaver nephrectomy was the largest avoidable loss of potentially transplantable kidneys. A similar but unavoidable loss occurred through sudden death of the possible donor.

University Department of Surgery and Nuffield Transplantation Surgery Unit, Western General Hospital, Edinburgh
A McL JENKINS, MB, FRCS, lecturer

Introduction

The shortage of cadaver kidneys for transplantation is a well-recognised problem in most transplant centres. Probably many potentially satisfactory cadaver kidneys are lost mainly through failure to refer the possible donors to a transplant unit. Experience of the Nuffield Transplantation Surgery Unit (NTSU) at the Western General Hospital (WGH), Edinburgh, however, has shown that we failed to harvest kidneys from a high proportion (77%) of possible donors who were referred to the unit. To assess the incidence and reasons for this failure we surveyed all