0016-5107/95/4206-0555$5.00 + 0
GASTROINTESTINAL ENDOSCOPY
Copyright © 1995 by the American Society for Gastrointestinal Endoscopy

Descriptive features of gastric ulcers: do endoscopists agree on what they see?

Peter W. Moorman, MD, Peter D. Siersema, MD, PhD, Astrid M. van Ginneken, MD, PhD

Rotterdam, The Netherlands

Background: **Little is known about the interobserver variation between endoscopists on descriptive morphologic features.**
Methods: **This study describes the agreement among 10 endoscopists on their description of 12 morphologic features, using 10 photographs of gastric ulcers, and on their eventual interpretation. The endoscopists used a form with predefined options for description.**
Results: **Kappa value was on average 0.36 for descriptive features and 0.31 for interpretation. The proportion of endoscopists agreeing on descriptive features was on average 84%, and 81% on interpretations. The chance of an endoscopist describing all 12 morphologic features of an ulcer on a photograph exactly the same as a colleague ranged from 4% to 46% (average 15%). A positive correlation between agreement in description and interpretation (0.75, $p < 0.05$) was found.**
Conclusions: **These results indicate a poor agreement between endoscopists in their translation of visual observations into descriptive terms. The positive correlation between agreement in description and interpretation suggests disagreement in description as an important cause for disagreement in interpretation. We believe that the use of more explicit descriptive terms will improve agreement in description and in subsequent interpretation. (Gastrointest Endosc 1995;42:555-9.)**

Reporting plays an essential role in endoscopy. Endoscopy reports convey the findings of an endoscopic examination to the physician who requested it and serve as reference material for future examinations. The importance of reports is underlined by the development of guidelines for the contents[1, 2] and the development of a standardized terminology.[3]

Describing the same topics and using the same terminology, however, is no guarantee that endoscopists will describe identical findings in a similar way. Previous research has shown that endoscopic findings lack accuracy (e.g., size estimates suffer from an underestimation of up to 30%[4, 5] and that essential parts of endoscopy reports such as diagnoses and interpretations suffer from interobserver variation.[6-8] Interobserver variation is not unique to endoscopy. It is also reported in other clinical disciplines[9-10] and histologic reporting.[11] Most endoscopists will agree that an endoscopic diagnosis such as "malignant gastric ulcer" is unreliable because of a lack of accuracy (subsequent histologic diagnosis may prove otherwise)[12] and reproducibility (another endoscopist may classify the same lesion as benign).

In contrast, little is known about the reliability of descriptive statements such as "the ulcer has an irregular border." Insight into such reliability is important because it provides information on (1) the probability that two endoscopists describe morphologic descriptive features the same way, and (2) the a priori predictive value of a feature for the diagnosis of a lesion (e.g., when a sharply demarcated ulcer edge

Received June 28, 1994. For revision December 14, 1994. Accepted March 1, 1995.

From the Department of Medical Informatics and the Department of Internal Medicine, Erasmus University, Rotterdam, The Netherlands.

Financial support for this research provided by Glaxo BV, Zeist, The Netherlands.

Reprint requests: P.W. Moorman, MD, Medical Informatics, Medical Faculty Ee 2110, Erasmus University, PO Box 1738, 3000 DR Rotterdam, The Netherlands.

37/1/66085

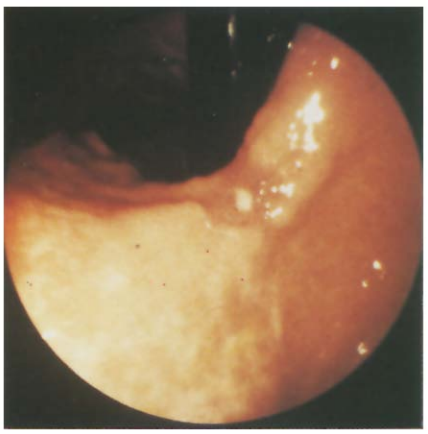

Figure 1. Example of one of the photographs used.

is present in 11 of 20 benign ulcers and in 2 of 20 malignant ulcers,[13] accurate assessment of the predictive value of such statements is not possible when the reliability of a description of the ulcer demarcation is not known).

Determining the reliability of descriptive features in respect to the truth is difficult, because defining the gold standard is virtually impossible. Microscopy, for example, cannot state with more certainty whether the base of an ulcer is regular or irregular. Another measure of reliability, however, is the agreement between endoscopists on descriptive statements. In this study, we assessed the interobserver variation between endoscopists regarding descriptive morphologic features and interpretation of what is considered a difficult endoscopic diagnosis: gastric ulcer.

MATERIALS AND METHODS

Photographs, endoscopists, and evaluation form

From our Gastroenterology Unit, we retrospectively obtained the 10 most recent slides of gastric ulcers that were of reasonable technical quality (i.e., sharpness, contrast). We made paper-prints (photographs) of these slides. An example of one of the photographs used in this study is shown in Figure 1. Ten experienced endoscopists were asked to participate, and all agreed. Two endoscopists work in our university hospital, the remaining eight practice in hospitals affiliated with the university.

We asked the endoscopists to evaluate the 10 photographs using a specially designed evaluation form. The form offered predefined options to describe 12 main morphologic features of a gastric ulcer. To give an example: the feature shape could be described by the options circular, oval, linear, serpiginous, and irregular. The endoscopists were allowed to select more than one option per feature. For each photograph and for each feature, the endoscopist had the option to indicate that a reliable description of that particular feature was not possible. When an endoscopist indicated that bleeding stigmata were present, he could specify these by selecting one or more of the following: clot, visible vessel, or active bleeding.

After describing an ulcer, the endoscopists were asked to give a diagnostic impression of that ulcer, using a 5-point scale ranging from possibly benign to possibly malignant. We will refer to this diagnostic impression as *interpretation*.

Data analysis

To analyze agreement, we grouped the descriptions given by the endoscopists into three categories (Table 1). The first category comprises descriptions that can be regarded as being contradictory to those in the second category. The third category constitutes the answers in which the endoscopist indicated that no reliable description for that feature could be given. When an endoscopist had described a feature using options in both category I and II, the description was assigned to the category II.

Agreement can be expressed in several ways. In our study we used the following: kappa, proportion of agreeing endoscopists, and the chance that individual endoscopists produce the same description of a photograph (chance of same description).

Kappa. For agreement between two endoscopists on the three categories of an ulcer feature, Cohen's kappa[14-16] can be calculated by the formula:

$$\kappa = \frac{Po - Pe}{1 - Pe}$$

where Po is the observed agreement and Pe is the agreement expected by chance. (The calculation of Po and Pe is described in the appendix). The overall kappa value for all endoscopists may be calculated by averaging all pairwise calculated kappas. When Pe equals one, kappa cannot be calculated. Kappa can range from −1 to 1, and is constructed to be zero when the obtained agreement can be entirely attributed to chance. The interpretation of kappa values is somewhat subjective, but kappa values above 0.75 are considered to represent excellent agreement, and values below 0.40 poor agreement.[17] Although kappa is a generally accepted measure, a difficulty in the interpretation is that kappa is also affected by the presence of bias between observers (e.g., when endoscopists assign observations predominantly to one category) and by the distribution of data across the categories.[18]

Proportion of agreeing endoscopists. In most studies assessing interobserver agreement, only two or three observers are involved. The 10 observers in our study permit us to express agreement also in the proportion of agreeing endoscopists (PAE). As PAE signifies the chance that an endoscopist would describe a feature the same as a colleague, it is a more intuitive and illustrative measure of agreement than kappa. When the answers are dichotomized in contradictory statements, then the proportion of endoscopists agreeing on that topic is defined as:

$$\text{PAE} = \frac{100x}{(x + y)}\ \%$$

where x endoscopists (the largest proportion) state option X, and y endoscopists state Y. PAE ranges from 50% (half of the endoscopists state X, the other half state Y) to 100% (all endoscopists state X or Y). When, for example, PAE is 80%, then every fifth endoscopist has stated the contrary of the

Table 1.
The morphologic features of gastric ulcers, the options for description on the evaluation form, and the categories to which they were assigned

Subject	Options		
	Category I	Category II	III
Morphologic features			
Shape	Circular, oval, linear	Serpiginous, irregular	Npd
Depth	Superficial	Medium deep, deep	Npd
Base: regularity	Regular	Irregular	Npd
Base: exudate	Absent	Present	Npd
Border: elevation	Flat	Partially raised, completely raised	Npd
Border: regularity	Regular	Irregular	Npd
Border: undermining	Absent	Present	Npd
Surrounding mucosa: color	Normal	Red, pale	Npd
Surrounding mucosa: swelling	Absent	Present	Npd
Surrounding mucosa: nodules	Absent	Present	Npd
Demarcation from surroundings	Sharp	Vague	Npd
Stigmata of bleeding	Absent	Present	Npd
Diagnostic impression			
Interpretation (from possibly benign to possibly malignant)	1, 2	4, 5	3

Npd, Not possible to give a reliable description.

other four. Mean PAE for a feature was calculated by averaging the PAE of that feature on every photograph.

In this study, differences in the values of kappa and PAE originate from the fact that (1) unlike PAE, the attribution of chance agreement is eliminated in the calculation of kappa, and (2) kappa also includes disagreement among endoscopists on whether or not a reliable assessment could be given.

Chance of same description (CSD). When we assume that PAEs for the features of a given photograph are independent, then we can calculate the chance that, given a description, a second endoscopist would give exactly the same description for that photograph by the formula:

$$CSD = \prod_{i=1}^{N} PAEi$$

where N is the number of features, in our case 12. Note that the assumption of independence does not relate to independence of the appearance of features, but to independence of agreement. We think that such an assumption is acceptable, although we acknowledge the fact that some degree of correlation between the agreement on various features may exist.

In addition, we calculated Spearman rank correlation coefficient to test whether agreement on interpretation ($PAE_{\text{interpretation of photo x}}$) correlates with agreement on description ($CSD_{\text{photo x}}$).

RESULTS

Descriptive features. Kappa values for descriptive features ranged from 0.06 to 0.59, and averaged 0.36 (Table 2). Highest kappa value was found for the feature describing whether or not the ulcer was superficial. Poor agreement among the endoscopists was found for the description of the shape (regular vs irregular), elevation of the border, and all features concerning the surrounding mucosa.

Table 2.
Kappa value and proportion of agreeing endoscopists on descriptive gastric ulcer features and interpretation

Feature	Kappa	PAE	
		%	Range
Shape	0.38	86	60-100
Depth	0.59	88	70-100
Base: regularity	0.40	82	60-100
Base: exudate	*	99	90-100
Border: elevation	0.33	78	60-100
Border: regularity	0.41	81	55-100
Border: undermining	0.46	91	60-100
Surrounding mucosa: color	0.23	79	50-100
Surrounding mucosa: swelling	0.06	69	50-87
Surrounding mucosa: nodules	0.20	81	50-100
Demarcation from surroundings	0.44	87	60-100
Stigmata of bleeding	0.43	88	57-100
Average	0.36	84	
Interpretation	0.31	81	55-100

*Kappa could not be calculated (Pe = 1).

The proportion of agreeing endoscopists (PAE) on gastric ulcer features ranged from 69% to 99%, and was on average 84% (Table 2). For the features, *presence of exudate* and *undermining of the border*, the PAE was larger than 90%. PAE was less than 80% for the features *flat or elevated* border, *normal or abnormal color* of surrounding mucosa, and *presence of swelling* of surrounding mucosa. For the absence or presence of *active bleeding, a clot* and *a visible vessel*, PAEs respectively were 95%, 92%, and 79%.

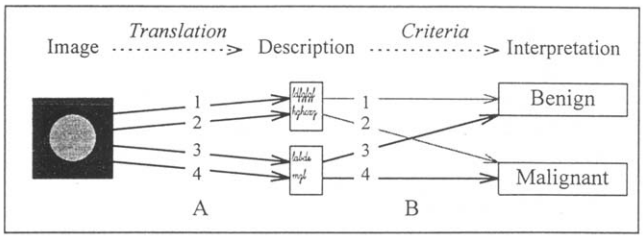

Figure 2. Two ways of coming to different interpretations. Endoscopists 1 and 2 give the same description of the lesion, but this description differs from the one given by endoscopists 3 and 4 **(A).** Starting from the same description, endoscopist 1 and 2 apply different criteria, and therefore arrive at a different interpretation **(B).**

Interpretations. The average proportion of endoscopists agreeing on the interpretation of a photograph of a gastric ulcer was 81%. Kappa value for the interpretation was also slightly below the average of descriptive features, namely 0.31.

Chance of same description (CSD). The chance of an endoscopist describing all 12 morphological features of an ulcer on a photograph exactly the same as a colleague (CSD) ranged from 4% to 46% (average 15%).

Correlation between description and interpretation. Spearman correlation between CSD of a photograph and the average PAE on the interpretation was 0.75 ($p < 0.05$).

DISCUSSION

This study describes the agreement among 10 endoscopists on their description of 12 morphologic features using photographs of gastric ulcers and on their eventual interpretation. The average kappa value for interpretation of gastric ulcers (0.31) indicates poor agreement between endoscopists, and reflects that interpretation is a complex process. Other studies[6-8] have already shown an average to low agreement on endoscopic diagnoses, but do not permit any conclusion on the nature of the low agreement. Two hypotheses may account for the lack of agreement on interpretations. First, endoscopists may differ in their criteria about what constitutes a malignant or benign ulcer (Fig. 2B). Second, they agree in their criteria, but fail to translate their visual observation in equal descriptive terms (Fig. 2A). If this second hypothesis dominates, endoscopists produce different descriptions on the basis of a given image of a lesion, but would arrive at the same interpretation on the basis of a given description of a lesion.

The correlation we found between agreement on description and agreement on interpretation (0.75, $p < 0.05$) supports the second hypothesis. This correlation indicates that where endoscopists give the same description, they also tend to arrive at the same interpretation; while giving different descriptions, their interpretations also differ. Disagreement in description thus accounts for low agreement on interpretation.

Disagreement in description may play an important role in daily practice: it is comparably low to agreement on interpretation (0.36 vs 0.31). The proportion of endoscopists agreeing on a descriptive feature was on average 84%, which signifies that if 20 endoscopists were to assess a single feature, then 17 would state regular and 3 irregular, or vice versa. On the assumption of independence of agreement on features, the chance that two endoscopists describe all 12 features in the same way (CSD) becomes very small, 15%.

These are important observations, as it also calls for caution in the interpretation of relations between descriptive endoscopic features and other observations, e.g., histological diagnoses. It is not inconceivable that the poor correlation between, for example, an irregular base and a (histologically) malignant ulcer is largely due to endoscopists failing to agree what constitutes an irregular base. In general, clinical studies relying on descriptive morphologic features in endoscopy (such as comparing effects of medications) pay little or no attention to the potential interobserver variation at the observational level.

How well do the discussed measures of agreement reflect reality? As compared with clinical practice, factors that may have caused underestimation and overestimation of kappa, PAE, and CSD need to be considered.

Underestimation of agreement. Because photographing gastric ulcers is not a standard procedure, photographs of interesting or difficult ulcers may be overrepresented in this study. Furthermore, the endoscopists made their descriptions from two-dimensional photographs and did not actually perform endoscopies themselves, depriving them from looking at the ulcers from different angles (pseudo three-dimensional view). However, the endoscopists had the option to indicate that no reliable description of a feature could be given; an option that was not used very often. In addition, our data do not indicate that this limitation played an important role, as agreement on two-dimensional features did not differ much from agreement on three-dimensional features. In fact, it was surprising that the highest kappa value was obtained for the assessment of depth.

Overestimation of agreement. Agreement in real practice may even be lower than the agreement we found, as in our study the endoscopists were confronted with the fact of a present ulcer, and kappa for the identification of the presence of an ulcer has been reported to be only 0.7.[7] Furthermore, as we categorized the descriptions given, endoscopists may also disagree within the same category. For example, the category *abnormal color of the surrounding mucosa*

included the options red and pale. Disagreement within this category abnormal color was still 10%.

What could be done to improve agreement between endoscopists on descriptive features? One could start with making the meaning of terms explicit. Although a statement such as *irregular border* may seem unambiguous, we found it could have two meanings, namely the elevation is irregular in height, or the elevation is irregular in width. Thus, the same assessment is given in differing situations. Making the meaning of descriptive terms explicit is important, but does not necessarily require a pure linguistic approach. When endoscopists describe features, it is likely that they use conceptual reference images; they compare what they see with an image in their memory. Providing the endoscopists with equal descriptions of reference images may therefore already improve agreement in description. Although this option may be realizable in an educational setting, it seems impractical in the daily clinical setting unless we make use of computer tools. In the future, we may envision an endoscopic reporting program in which, on selecting a term, the user is provided with images in which that term is visually represented.

We believe that reducing disagreement in endoscopic descriptions will increase the value of endoscopy reports in practice and research. Meanwhile, it seems important that clinical studies should strive to formulate descriptive features as explicitly as possible and should state the number of endoscopists that have participated in the study. For clinical practice, we believe that adding a photograph of the observed lesion to the report will help to ensure that the correct message is conveyed to the referring physician and to the endoscopist performing follow-up examinations.

ACKNOWLEDGMENT

We thank Dr. M. van Blankenstein and Dr. J. van der Lei for their help in the preparation of this document.

REFERENCES

1. American Society for Gastrointestinal Endoscopy. Quality assurance of gastrointestinal endoscopy: an information resource manual. Manchester, Massachusetts: ASGE, 1989.
2. Schapiro M. Computerization of endoscopic reports—an ASGE proposal. Endoscopy 1992;2(Suppl):478-80.
3. Maratka A. Terminology, definitions and diagnostic criteria in digestive endoscopy. 2nd ed. Omed database of digestive endoscopy. Bad Homburg: Normed Verlag, 1989.
4. Fennerty MB, Davidson J, Emerson SS, et al. Are endoscopic measurements of colonic polyps reliable? Am J Gastroenterol 1993;88:496-500.
5. Margulies C, Krevsky B, Catalano MF. How accurate are endoscopic estimates of size? Gastrointest Endosc 1994;40: 174-7.
6. Bendtsen F, Skovgaard LT, Sørensen TIA, Matzen P. Agreement among multiple observers on endoscopic diagnosis of esophageal varices before bleeding. Hepatology 1990;11:341-6.
7. Bytzer P, Havelund T, Møller Hansen J. Interobserver variation in the endoscopic diagnosis of reflux esophagitis. Scand J Gastroenterol 1993;28:119-25.
8. Woolf GM, Riddell RH, Irvine EJ, Hunt RH. A study to examine agreement between endoscopy and histology for the diagnosis of columnar lined (Barrett's) esophagus. Gastrointest Endosc 1989;35:541-4.
9. Koran LM. The reliability of clinical methods, data and judgments. N Engl J Med 1975;293:642-6, 695-701.
10. Komaroff AL. The variability and inaccuracy of medical data. Proc IEEE 1979;28:1196-207.
11. Christensen AH, Gjørup T, Hilden J, et al. Observer homogeneity in the histologic diagnosis of *Helicobacter pylori*. Scand J Gastroenterol 1992;27:933-9.
12. Dekker W, Tytgat GNJ. Diagnostic accuracy of fiberendoscopy in the detection of upper intestinal malignancy. A follow-up analysis. Gastroenterology 1977;73:710-4.
13. Gabrielsson N. Benign and malignant gastric ulcers. Evaluation of the different diagnostics in roentgen examination and endoscopy. Endoscopy 1972;4:73-83.
14. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20:37-46.
15. Fleiss JL. Statistical methods for rates and proportions. New York: J. Wiley & Sons, 1981.
16. Schouten HJA. Statistical measurements of interobserver agreement [Thesis]. Rotterdam: Erasmus University, 1985.
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:671-9.
18. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423-9.

APPENDIX

Suppose two raters classify N subjects as belonging to one of three categories. The result can be arranged in a 3 × 3 table as follows:

		Observer 1			
		I	II	III	Total
Observer 2	I	a	b	c	j_1
	II	d	e	f	j_2
	III	g	h	i	j_3
	Total	k_1	k_2	k_3	N

where *d* is the number of subjects assigned to category I by observer 1, and to category II by observer 2. The proportion of observed agreement is $Po = (a + e + i)/N$. Agreement expected by chance is $Pe = (j_1k_1 + j_2k_2 + j_3k_3)/N^2$.