Scandinavian Journal of Gastroenterology, 2007; 42: 1106–1112

informa
healthcare

ORIGINAL ARTICLE

Impact of observer variability on the usefulness of endoscopic images for the documentation of upper gastrointestinal endoscopy

ANNE METTE ASFELDT[1,3], BJØRN STRAUME[1] & EYVIND J. PAULSSEN[2,3]

[1]*Institute of Community Medicine,* [2]*Institute of Clinical Medicine, University of Tromsø, Norway, and* [3]*Department of Gastroenterology, University Hospital of North Norway, Tromsø, Norway*

Abstract

Objective. Endoscopy is an observer-dependent diagnostic method, which, until recently, has lacked precise guidelines for written reports. There is an increasing demand for improvement in endoscopy records, which may necessitate the supplementation of image documentation. The aim of this study was to estimate interobserver as well as intra-observer variability in the assessment of images from gastroscopy. ***Material and methods.*** We designed an Internet interface presenting endoscopy images, accompanied by a multiple-choice questionnaire for assessing pathology in the images. Ten images from the distal oesophagus and 10 images from the pyloric antrum were chosen. In order to study interobserver variability, physicians with varying endoscopy experience were invited to complete the questionnaire. The physicians were re-invited 5 months later to assess the same images again, this time in order to assess intra-observer variability. Kappa statistics were used for analysis of agreement. ***Results.*** Initially, 13 of 20 invited physicians responded. Interobserver agreement varied between poor ($\kappa < 0.2$) and moderate ($0.4 < \kappa < 0.6$). In the second part of the study, 10 of 11 invited physicians responded. Intra-observer agreement varied between moderate ($0.4 < \kappa < 0.6$) and good ($0.6 < \kappa < 0.8$). A higher level of experience does not imply either better interobserver or better intra-observer agreement. Images of concise endoscopy findings, such as the presence of an ulcer, resulted in better agreement than did the assessment of images of less definable findings. ***Conclusion.*** The variability in the interpretation of endoscopy images is large. We therefore believe that systematic inclusion of a set of images into endoscopy reports will improve their quality.

Key Words: *Documentation, endoscopy, epidemiology, gastrointestinal, gastroscopy, image, observer variation*

Introduction

Endoscopy as a diagnostic method developed along with the use of fibre-optic endoscopes at a time when systematic documentation of images was not readily available. In Norway, and most other Western countries, image documentation in gastrointestinal (GI) endoscopy has focused on the pathological findings. This is in contrast to the situation in Japan, where systematic image documentation of the endoscopic procedure is more widely used [1]. Today, the possibility of digital storing of images offers far better means of obtaining and keeping such documentation.

There is an increasing demand for proper documentation of performed procedures. Standardization of the endoscopy record has long been recognized as a possible means of improving documentation in endoscopy, initially fronted by Z. Mařatka [2], further emphasized by the European Society of Gastrointestinal Endoscopy (ESGE) in developing the minimal standard terminology for digestive endoscopy (MST) [3] which, in turn, was adopted by the World Organization of Gastrointestinal Endoscopy (OMED). As a further attempt to improve documentation, the ESGE has presented additional guidelines for standardized image documentation in upper and lower GI endoscopy [1].

Clinical disagreement is a well-known challenge in most, if not all, fields of medicine. It is important to address disagreement in order to find ways of minimizing it. Clinical disagreement in gastroscopy

Correspondence: Anne Mette Asfeldt, MD, Institute of Community Medicine, University of Tromsø, NO-9037 Tromsø, Norway. Tel: +47 7 7644 885. Fax: +47 7 7644 831. E-mail: anne.mette.asfeldt@ism.uit.no

(Received 14 November 2006; accepted 3 February 2007)

ISSN 0036-5521 print/ISSN 1502-7708 online
DOI: 10.1080/00365520701259240

can be illustrated by studying the observer variability. Several studies in the field of gastroenterology have addressed observer variability in more site-specific pathology. Various grading systems for oesophagitis have been studied, e.g. by Rath et al. who compared interobserver agreement for the Los Angeles (LA) classification of erosive oesophagitis, the MUSE scoring system and the Savary-Miller system, and found the first two to be the most reliable. They also found a higher degree of agreement among the more experienced endoscopists [4]. Contrary to what one might expect, in their validation study of the LA classification, Lundell et al. found that greater experience did not result in a higher degree of agreement [5].

Bendtsen et al. studied the diagnosis of oesophageal varices and found considerable variation among both experienced and less experienced endoscopists [6]. Several studies have investigated observer variability in colonoscopy. de Lange et al. concluded that the interobserver agreement in assessing ulcerative colitis is satisfactory among trained endoscopists [7], although with a potential for improvement by standardization of the text report [8]. Orlandi et al. arrived at a similar conclusion [9]. What these studies have in common is the use of specific grading systems in the assessment of the severity of inflammatory bowel disease.

Even though upper endoscopy is a well-established procedure, we believe that a study on variability in the assessment of images from upper endoscopy is of interest, especially in the light of the increasing availability of digital recording of images and thus the improved possibility of supplementing the endoscopy record. With this study, our aim was to evaluate observer variability in gastroscopy, in order to assess the usefulness of images for endoscopy documentation.

Materials and methods

Endoscopy examinations

All examinations were performed in spring 2004 as a part of the Sørreisa II study; a population-based study of GI disorders in the municipality of Sørreisa in Norway. Endoscopy examinations were carried out using Olympus GIF-160 video gastroscopes. As part of the study, a standard set of images from each individual were made on a routine basis and stored using the Endobase III software (Microsoft Corporation, Redmond, Wash., USA). The protocol for images was chosen for research purposes and implied one image from the following sites; distal oesophagus, pyloric antrum, gastric fundus and duodenal bulb.

Selection of images

The images were selected to ensure their technical quality and to reflect a broad spectrum of pathology common to most endoscopists. There were few images of normal mucosa or severe pathology. Ten images from the lower third of the oesophagus, and 10 images from the pyloric antrum were chosen from a total of 19 individuals. The images showed normal oesophageal or gastric mucosa, as well as other typical findings such as gastric ulcer and erosive prepyloric changes [10].

Selection of respondents

Gastroenterologists, or physicians who were known to perform endoscopy examinations, working at public hospitals in either a general internal medicine department or a department of gastroenterology were considered for participation. For practical reasons, we limited the invitation to physicians for whom contact information was readily available. Twenty physicians, mainly in our health region of Northern Norway, were invited to participate in the study by e-mail, and after one month non-responders were reminded of the study by e-mail.

Presentation and assessment of images

The images were presented in Joint photographic Experts Group (JPEG) format 450×464 pixels on an Internet interface, accessible only to invited physicians. No patient data were available, nor were patients' characteristics or symptoms. The physicians had unlimited time within the study period to view the pictures online. The respondents also provided information on endoscopy experience. Data were collected online and entered directly into a database at the University of Tromsø. The Internet interface consisted of 10 images from the distal oesophagus and 10 images from the pyloric antrum, together with a multiple-choice questionnaire containing a question regarding endoscopy experience, as well as three questions for every oesophageal image, and five questions for every antral image. The questions reflected a simplified version of the MST, which includes the LA classification for oesophagitis. Our interest in this study was the variability between and within observers in the assessment of images, not whether they reached a correct diagnosis. Thus, the assessments were not measured against a "gold standard".

In the invitation, we pointed out that a standardized assessment was chosen, and we presumed that the respondents were familiar with the LA classification. In addition, the respondents were not given any

guidelines on answering the questionnaire. The questions are presented in Table I.

An anonymous response was an option, but the respondents were invited to identify themselves in order to facilitate a second contact for evaluation of intraobserver variability.

Thirteen physicians responded, and the first part of the study, dealing with interobserver variability, is based on their assessments. Eleven of the 13 respondents identified themselves, and 5 months later they were invited again to assess the same images. Ten of them responded, and the second part of the study, on intra-observer variability, is based on these 10 pairwise assessments.

Figures 1 and 2 are examples of images from the oesophagus and the pyloric antrum, respectively.

Statistical analysis

Interobserver variability was analysed from the initial assessment from all 78 possible pairs of respondents of each of 10 images for every posed question, leading to 780 pairwise assessments. First, the analysis was carried out for all responders as a whole, after which a subanalysis was done by dividing the respondents into two groups based on experience: "highly experienced" (>1000 upper endoscopies performed) and "moderately experienced" (200–1000 upper endoscopies performed). Likewise, intraobserver variability was estimated from the image assessments for responders who participated in both parts of the study.

The variability of categorical data was measured using kappa statistics [11], whereas a weighted kappa was used for analysis of ordinal data (oesophagitis) [12]. Agreement, based on the value of kappa, was categorized, as described by Altman, as poor ($\kappa \leq 0.2$), fair ($0.21 \leq \kappa \leq 0.40$), moderate ($0.41 \leq \kappa \leq 0.60$), good ($0.61 \leq \kappa \leq 0.80$) or excellent ($0.81 \leq \kappa \leq 1.00$) [12]. The precision of kappa was measured by its 95% confidence interval (CI). The analysis was done using SPSS statistical software (SPSS Inc., Chicago, Ill., USA) for cross-tabulation of results and using Excel software (Microsoft Corporation) for measures of kappa and confidence intervals.

Table I. Questionnaire as presented on the Internet site.

Subject	Question	Options
Experience	How many gastroscopies have you performed?	Less than 200 200–1000 >1000
Oesophageal images [10]	Oesophagitis according to the LA classification?	None/A/B/C/D
	Suspected metaplasia?	Yes/No
	Hiatus hernia?	Yes/No/Uncertain
Gastric images [10]	Normal mucosa?	Yes/No
	Oedematous mucosa?	Yes/No
	Erythematous mucosa?	Yes/No
	Erosion(s)?	Yes/No
	Ulcer(s)?	Yes/No

Each respondent assessed 20 images by marking one option to each question related to the images.

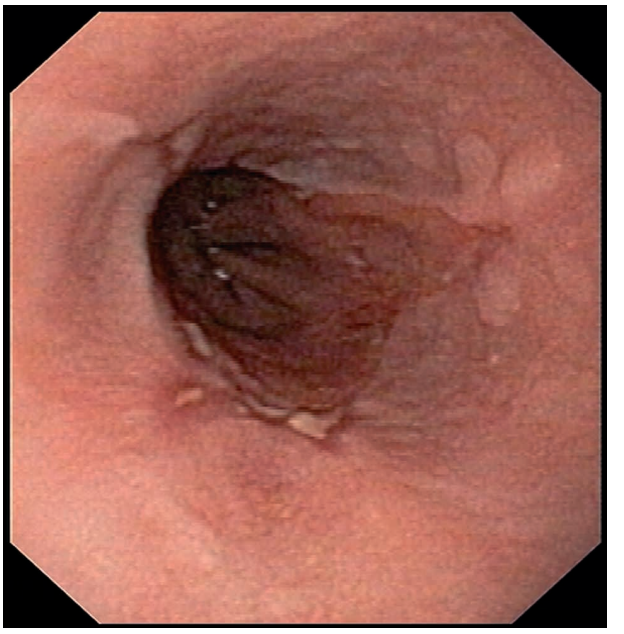

Figure 1. Endoscopic view of the distal part of the oesophagus.

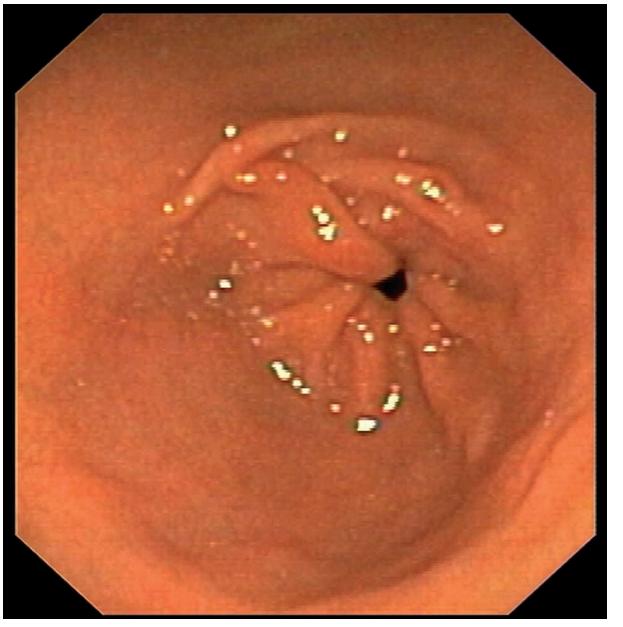

Figure 2. Endoscopic view of the antrum.

Ethics

The endoscopy images were obtained from participants in an epidemiology survey, which was approved by the local Regional Committee for Medical Research Ethics. The images were presented without personal identification.

Results

Interobserver variability

Thirteen out of 20 invited physicians (65%) responded to the survey. In the assessment of images from the oesophagus, there was full agreement on 3 out of 30 questions; i.e. 1 question on the presence of hiatus hernia and 2 on the presence of metaplastic changes. None of the images of oesophagitis was assessed with full agreement from all the respondents. On the contrary, three images were assessed using all five available categories of grading. Figure 1 is an example of an image from the oesophagus in which oesophagitis was assessed with LA classification grades A, B, C and D, as well as "not present".

In the assessment of images from the pyloric antrum, all 13 respondents agreed fully on 13 out of 50 questions. For one antral image, (Figure 2), the respondents agreed fully on all five questions.

The interobserver agreement is summarized in Table II. Level of agreement is defined as described in the Materials and methods section.

In the analysis of data from all 13 respondents (78 possible pairs), three questions held moderate agreement. These questions were on rather well-defined findings, i.e. normal gastric mucosa, gastric erosions and ulcers. The remaining questions held fair or poor agreement, with the question on erythema in the gastric mucosa being the most difficult to agree upon.

When the data were divided according to "highly experienced" and "moderately experienced" endoscopists, we found that higher experience was not followed by a higher level of agreement. On the contrary, the "highly experienced" endoscopists held poor agreement on two questions, and less agreement than expected by chance in the question on erythema in the gastric mucosa.

Intra-observer variability

Five months later, 10 out of 11 participants responded to an invitation to assess the images again. A summary of the intra-observer variability is presented in Table III.

In the assessments of all 10 respondents, agreement varied between moderate and good. When divided into groups based on experience, we found

Table II. Inter-observer variability in the assessment of images from gastroscopy.

	All 13 respondents (78 pairs)			Five highly experienced respondents (10 pairs)			Eight moderately experienced respondents (28 pairs)		
Question on	Observed agreement	Kappa (95% CI)	Agreement	Observed agreement	Kappa (95% CI)	Agreement	Observed agreement	Kappa (95% CI)	Agreement
Oesophagitis*	0.75	0.25 (0.16–0.34)	Fair	0.69	0.19 (−0.26 to 0.63)	Poor	0.79	0.28 (0.12–0.45)	Fair
Oesophageal metaplasia	0.70	0.40 (0.34–0.47)	Fair	0.64	0.23 (0.10–0.47)	Fair	0.73	0.45 (0.35–0.56)	Moderate
Hiatus hernia	0.51	0.25 (0.19–0.30)	Fair	0.56	0.26 (0.10–0.43)	Fair	0.50	0.25 (0.16–0.34)	Fair
Normal gastric mucosa	0.84	0.47 (0.39–0.56)	Moderate	0.76	0.35 (0.12–0.57)	Fair	0.88	0.57 (0.43–0.71)	Moderate
Oedematous gastric mucosa	0.65	0.28 (0.21–0.35)	Fair	0.66	0.26 (0.06–0.46)	Fair	0.63	0.25 (0.13–0.36)	Fair
Erythematous gastric mucosa	0.59	0.18 (0.11–0.25)	Poor	0.46	−0.08 (−0.28 to 0.11)	Poor	0.65	0.26 (0.14–0.37)	Fair
Gastric erosion(s)	0.72	0.41 (0.34–0.47)	Moderate	0.80	0.59 (0.43–0.75)	Moderate	0.67	0.28 (0.16–0.40)	Fair
Gastric ulcer(s)	0.85	0.50 (0.42–0.58)	Moderate	0.84	0.41 (0.15–0.68)	Moderate	0.85	0.54 (0.41–0.67)	Moderate

*Weighted agreement and weighted kappa (results from pairs of respondents with one "highly experienced" and one "moderately experienced" endoscopist are not presented, which explains why – on the question of oedematous gastric mucosa – the kappa value for all respondents is higher than that for both "moderately" and "highly" experienced respondents).

Table III. Intraobserver variability in the assessment of images from gastroscopy.

	All 10 respondents (10 pairs)			Five highly experienced respondents (5 pairs)			Five moderately experienced respondents (5 pairs)		
Question on	Observed agreement	Kappa (95% CI)	Agreement	Observed agreement	Kappa (95% CI)	Agreement	Observed agreement	Kappa (95% CI)	Agreement
Oesophagitis*	0.81	0.43 (0.20–0.66)	Moderate	0.76	0.31 (−0.03 to 0.65)	Fair	0.86	0.55 (0.25–0.86)	Moderate
Oesophageal metaplasia	0.85	0.70 (0.55–0.84)	Good	0.76	0.52 (0.28–0.76)	Moderate	0.94	0.86 (0.74–1.01)	Excellent
Hiatus hernia	0.68	0.52 (0.38–0.65)	Moderate	0.64	0.44 (0.23–0.65)	Moderate	0.72	0.58 (0.39–0.77)	Moderate
Normal gastric mucosa	0.89	0.64 (0.43–0.84)	Good	0.86	0.58 (0.29–0.87)	Moderate	0.92	0.70 (0.42–0.98)	Good
Oedematous gastric mucosa	0.78	0.53 (0.35–0.70)	Moderate	0.78	0.54 (0.30–0.78)	Moderate	0.78	0.51 (0.26–0.77)	Moderate
Erythematous gastric mucosa	0.74	0.46 (0.28–0.64)	Moderate	0.66	0.32 (0.05–0.58)	Fair	0.82	0.59 (0.35–0.83)	Moderate
Gastric erosion(s)	0.80	0.59 (0.44–0.75)	Moderate	0.82	0.64 (0.42–0.85)	Good	0.78	0.55 (0.32–0.79)	Moderate
Gastric ulcer(s)	0.93	0.66 (0.41–0.90)	Good	0.90	0.49 (0.07–0.91)	Moderate	0.96	0.81 (0.56–1.07)	Excellent

*Weighted agreement and weighted kappa.

that on two questions, the five "moderately experienced" endoscopists held excellent agreement, and not lower than moderate agreement in any other question. The five "highly experienced" endoscopists did not obtain excellent agreement on any question. On the contrary, responses to the two questions regarding oesophagitis and gastric erythema held only fair agreement.

Discussion

We find that variability is extensive in the assessment of images from upper **e**ndoscopy. Similar findings are known from other diagnostic disciplines, e.g. interpretation of mammograms [13], diagnosis of vertebral fractures [14] and assessment of carotid plaques [15]. The interobserver variability in our study ranged from poor to moderate, with the highest level of agreement in response to questions regarding characteristic findings such as ulcer(s) or erosion(s). Even a widely used and well-evaluated classification system, such as the LA classification of oesophagitis [4,5], renders considerable variability in the reported assessments.

In the intra-observer part of this study, agreement was higher than that in the interobserver part, as would be expected. Even though endoscopists are more likely to agree with themselves than with each other, only two of the questions obtained an agreement level of "excellent" in the intra-observer study, this being in the group of "moderately experienced" endoscopists.

We had expected to find a higher level of agreement among "highly experienced" endoscopist than among the "moderately experienced". Other studies, however, do not fully concur with this assumption. Some studies [9,16,17] find that experience leads to a higher degree of agreement, whereas others do not [5,6].

Still images from gastroscopy fall short in documenting motility and other dynamic factors, which are equally as important as mucosal changes. Recording the whole examination on video could remedy such shortcomings, but for practical reasons it is uncertain whether video recordings represent a realistic means of routinely documenting gastroscopy. A standardized set of still images will always be second best, yet by far the more practical method. The ESGE has suggested a series of eight reference images for the documentation of upper endoscopic procedures [1].

In previous studies it has been highlighted that structured reporting can improve the quality of the endoscopy record [8,18]. Most classification systems used in endoscopy are, however, restricted to the description of certain features, e.g. oesophagitis. In

contrast, the MST [3] facilitates a complete standardized endoscopy record. The results from our study support the importance of such structured reporting.

As argued above, interpretation of endoscopy images is not an exact procedure. Accordingly, image documentation is important in order to reveal the ambiguity of gastroscopy. Without image documentation the written endoscopy record stands alone, with conclusions that may be more definite than are justified. The present possibilities of digital recording and distribution of images from endoscopy offer not only a better clinical practice, but also the possibility of better education and quality assessments of endoscopy as a diagnostic method. This has been pointed out by de Lange et al. in their study of an Internet interface for the assessment of endoscopy images [19].

The strength of our study is that it deals with several aspects of upper endoscopy simultaneously, and thus reflects daily clinical practice more than do most previous studies on the variability in endoscopy. In addition, our study covers both inter- and intra-observer variability, in an Internet interface, thus reflecting both the possibilities and limitations of digital imaging and documentation in endoscopy.

There are, of course, some limitations to this study. When first contacted, the respondents were informed that we planned an interobserver study as well as an intra-observer study. This could affect the intra-observer responses, resulting in an unfounded high degree of agreement. Despite this limitation, we do not believe that it alters the conclusions of this study. On the contrary, if the levels of agreement we observed are artificially high, this would only strengthen our conclusion. The non-responders in the study were not characterized in any particular manner, and we have no reason to believe that they should give rise to any bias.

The statistical method used in this study has some limitations. Kappa statistics are highly influenced by the prevalence of disease, and this limitation implies that kappa is purely descriptive of the agreement in the study in question. Kappa statistics do not imply the testing of hypothesis or estimation of "true" agreement. Therefore, there is no single answer as to how many responders or questions should be included in the study. With only two observers, which is often the case in agreement studies, extreme answers will have a great impact on kappa. None of the respondents in our study stood out with extreme responses, thus arguing for the validity of our data. With numerous questions to answer, the disagreement will stack up and lower the kappa values. There are other methods of agreement for analysis of continuous data, but with our data being nominal/ordinal, we chose to use kappa statistics despite their shortcomings. These are the considerations on which we have based our study design, and we believe that the numbers of respondents and questions are adequate to support the conclusions of the study. At the same time, we recognize that other designs are possible.

Despite these reservations, we conclude that there is considerable variability in the assessment of images from upper endoscopy. We argue that this uncertainty should be regarded in the structured endoscopy report by including a standardized set of images to document the findings. Further studies on the usefulness of image documentation in clinical practice are needed.

Acknowledgements

This project was supported with the aid of EXTRA funds from the Norwegian Foundation for Health and Rehabilitation, the National Association for Digestive Diseases, the Northern Norway Regional Health Authority and The University of Tromsø. We thank senior IT engineer Jarle Mathiasen for constructing the Internet site and the online questionnaire. We also thank those of our colleagues who contributed to this study by responding to the questionnaire.

References

[1] Rey JF, Lambert R. ESGE recommendations for quality control in gastrointestinal endoscopy: guidelines for image documentation in upper and lower GI endoscopy. Endoscopy 2001;33:901–3.

[2] Mařatka Z. Terminology, definitions and diagnostic criteria in digestive endoscopy. With the collaboration of the members of the Terminology Committee of the World Society of Digestive Endoscopy/OMED. Scand J Gastroenterol ; Suppl 1984;103:1–74.

[3] European Society of Gastrointestinal Endoscopy.. Minimal standard terminology in digestive endoscopy. Endoscopy 2000;32:162–88.

[4] Rath HC, Timmer A, Kunkel C, Endlicher E, Grossmann J, Hellerbrand C, et al. Comparison of interobserver agreement for different scoring systems for reflux esophagitis: impact of level of experience. Gastrointest Endosc 2004;60:44–9.

[5] Lundell LR, Dent J, Bennett JR, Blum AL, Armstrong D, Galmiche JP, et al. Endoscopic assessment of oesophagitis: clinical and functional correlates and further validation of the Los Angeles classification. Gut 1999;45:172–80.

[6] Bendtsen F, Skovgaard LT, Sorensen TI, Matzen P. Agreement among multiple observers on endoscopic diagnosis of esophageal varices before bleeding. Hepatology 1990;11:341–7.

[7] de Lange T, Larsen S, Aabakken L. Inter-observer agreement in the assessment of endoscopic findings in ulcerative colitis. BMC Gastroenterol 2004;4:9.

[8] de Lange T, Moum BA, Tholfsen JK, Larsen S, Aabakken L. Standardization and quality of endoscopy text reports in ulcerative colitis. Endoscopy 2003;35:835–40.

[9] Orlandi F, Brunelli E, Feliciangeli G, Svegliati-Baroni G, Di SA, Benedetti A, et al. Observer agreement in endoscopic assessment of ulcerative colitis. Ital J Gastroenterol Hepatol 1998;30:539–41.
[10] Nesland A, Berstad A. Erosive prepyloric changes in persons with and without dyspepsia. Scand J Gastroenterol 1985;20: 222–8.
[11] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measures 1960;20:37–46.
[12] Altman DG. Practical statistics for medical research. London: Chapman & Hall; 1991. pp 403–9.
[13] Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994;331:1493–9.
[14] Delmas PD, van de Langerijt L, Watts NB, Eastell R, Genant H, Grauer A, Cahall DL. Underdiagnosis of vertebral fractures is a worldwide problem: the IMPACT study. J Bone Miner Res 2005;20:557–63.
[15] Lovett JK, Gallagher PJ, Rothwell PM. Reproducibility of histological assessment of carotid plaque: implications for studies of carotid imaging. Cerebrovasc Dis 2004;18:117–23.
[16] Armstrong D, Bennett JR, Blum AL, Dent J, De Dombal FT, Galmiche JP, et al. The endoscopic assessment of esophagitis: a progress report on observer agreement. Gastroenterology 1996;111:85–92.
[17] Pandolfino JE, Vakil NB, Kahrilas PJ. Comparison of inter- and intraobserver consistency for grading of esophagitis by expert and trainee endoscopists. Gastrointest Endosc 2002; 56:639–43.
[18] Delvaux M, Crespi M, Armengol-Miro JR, Hagenmuller F, Teuffel W, Spencer KB, et al. Minimal standard terminology for digestive endoscopy: results of prospective testing and validation in the GASTER project. Endoscopy 2000;32: 345–55.
[19] de Lange T, Svensen AM, Larsen S, Aabakken L. The functionality and reliability of an Internet interface for assessments of endoscopic still images and video clips: distributed research in gastroenterology. Gastrointest Endosc 2006;63:445–52.