Observer reproducibility in grading dysplasia in colorectal adenomas: comparison between two different grading systems

C Fenger, M Bak, O Kronborg, H Svanholm

Abstract

The two most well known and well defined grading systems for dysplasia in colorectal adenomas were compared with regard to reproducibility. The Konishi-Morson system (KMS) operates with several histological and cytological variables and grades of mild, moderate, and severe dysplasia. The Kozuka system is based on the extent of nuclear pseudostratification and also has three grades of dysplasia (III-V). As the group of severe dysplasia is very large in this system, it was extended with two higher grades, similarly based on individual histological criteria, known hereafter as the extended Kozuka system (EKS). Fifty six adenomas were graded by two observers, each observer grading twice according to the KMS criteria and twice according to EKS criteria. Intraobserver reproducibility was excellent for the KMS and moderate for the EKS, but this was not significant. The overall interobserver reproducibility was similar (moderate) for the KMS and for the EKS. Kappa values for interobserver reproducibility on individual categories were excellent for severe dysplasia according to the KMS, but low for all other categories in both systems.

By simplifying both systems into two groups a high reproducibility can be obtained, but this implies that all the original grades (III–V) for the EKS must be grouped together. It is therefore recommended that a simplified KMS is used for further studies on the biological importance of dysplasia and for comparison between histological changes and other markers for colorectal neoplasia.

Odense University Hospital Odense, Denmark
Department of Pathology
C Fenger
M Bak
H Svanholm

Department of Gastrointestinal Surgery
O Kronborg

Correspondence to:
Dr C Fenger, Department of Pathology, Odense University Hospital, DK-5000, Odense, Denmark

Accepted for publication
6 December 1989

Adenomas are the most well known precursors of colorectal adenocarcinoma. Proper removal of an adenoma eliminates the malignant potential of that particular lesion, but the risk of developing new adenomas and possibly carcinomas, therefore, may be related to the number, location, size, architecture and grade of dysplasia in the removed adenomas.[1] Determination of the first three variables can be carried out objectively and it is generally agreed that adenomas should be grouped into tubular, tubulovillous, and villous according to the percentage of villous structure present—that is, less than 20%, 20–80%, and more than 80%, respectively. Grading of dysplasia can, however, be carried out according to different systems.[1–4]

The value of a grading system for adenomas depends on its ability to select patients at high risk of developing cancer, but also on its reproducibility. Only well defined systems are likely to fulfill both criteria. The two most well known and precisely described systems are those of Konishi and Morson[1] and Kozuka.[2]

The Konishi-Morson system[1] (KMS) is essentially a detailed description of the WHO system[4] and describes three grades—namely, mild, moderate, and severe dysplasia—using a combination of variables, including tubule configuration, nuclear polarity, orientation and structure, mucin content and location etc. The Kozuka system,[2] in contrast, describes five grades using a single variable—namely, the extent of epithelial pseudostratification, and the three highest grades (III–V) are regarded as true dysplasia. Kozuka's grade V will, however, inevitably include adenomas, which, according to the KMS, would be graded as moderate, and in our opinion therefore do not differentiate sufficient numbers of patients at possible high risk, a point of view also expressed by others.[5]

In the present study the Kozuka grade V was therefore subdivided into three grades. This system will be referred to as the "extended Kozuka system" (EKS). The term carcinoma in situ (CIS) was used where the glands showed a cribriform architecture and total loss of nuclear polarity. The term intramucosal carcinoma (IMC) was used for areas showing loosely scattered cells in the lamina propria, believed to represent early invasive growth. This is a subgroup of adenomas with a theoretical risk of metastases, as lymphatics have been shown to be present in the basal part of the lamina propria of the human colon.[6] The present study was carried out to test the intra- and interobserver reproducibility of the two systems.

Methods

Fifty six adenomas were selected by the surgeon (OK) from the files of the department of pathology. The selection was based exclusively on the original diagnosis and the only requirement was that all five grades of the EKS should be represented with several specimens. The original sections were used, all adenomas having been totally paraffin wax embedded and stained with haematoxylin and eosin. The number of sections from each adenoma varied from one to 34 with an average of five.

Table 1 Intraobserver agreement and reproducibility in grading 56 adenomas according to KMS

		MB run 1				*Agreement due to chance*		
		Mild	*Moderate*	*Severe*	*Agreement*		*Kappa*	*SE (κ)*
MB run 2:	Mild	12	1	0				
	Moderate	4	19	1	0·857	0·344	0·78	0·095
	Severe	0	2	17				
		CF run 1						
CF run 2:	Mild	24	2	0				
	Moderate	0	10	3	0·875	0·355	0·81	0·096
	Severe	0	2	15				

Table 2 Intraobserver agreement and reproducibility in grading 56 adenomas according to EKS

		MB run 1						*Agreement due to chance*		
		Mild	*Moderate*	*Severe*	*CIS*	*IMC*	*Agreement*		*Kappa*	*SE (κ)*
MB run 2:	Mild	3	0	0	0	0				
	Moderate	0	20	2	1	0				
	Severe	0	4	10	2	0	0·786	0·294	0·70	0·080
	CIS	0	0	1	9	0				
	IMC	0	0	1	1	2				
		CF run 1								
CF run 2:	Mild	6	4	0	0	0				
	Moderate	1	20	1	1	0				
	Severe	0	2	3	0	0	0·768	0·285	0·68	0·074
	CIS	0	1	0	10	0				
	IMC	0	0	0	3	4				

The adenomas were randomly renumbered and examined blind twice by each of the two pathologists (MB and CF), who had familiarised themselves with both systems. Each observer graded the adenomas twice according to the KMS and twice to the EKS, with an interval of at least one week. The highest grade of dysplasia was noted as the result.

The results were analysed with κ statistics[7 8] using a computer program developed in our department.[9] The term agreement was used for the overall or proportional number of cases given the same diagnosis among or within observers, including that part of the agreement which may have been due to chance. The term reproducibility (synonymous with repeatability) was used for that part of the agreement which may not have been explained by pure chance. An index of the reproducibility was given by the κ coefficient. The strength of reproducibility was expressed in terms of the arbitrary division of Svanholm *et al*,[9] according to which $\leq 0{\cdot}50$ is regarded as poor, $\kappa = 0{\cdot}51$–$0{\cdot}74$ as moderate, and $\geq 0{\cdot}75$ as excellent. The difference between κ values was tested according to the method of Cohen.[8]

Associations between categories within a given grading system were calculated according to Holman *et al.*[10] Thus if one of the pathologists assigned a random case to one category then the figures in the tables give the probability that the other pathologist would assign the same case to the same category or one of the other categories. Associations between categories of different grading systems were calculated according to Svanholm *et al.*[9]

Table 3 Overall interobserver agreement and κ statistic for two observers grading 56 cases of colonic adenomas on two separate occasions

	Agreement	*Agreement due to chance*	*Kappa*	*95% confidence limits of Kappa*
KMS:				
1 rating (n = 56)	0·714	0·334	0·57	0·38–0·76
2 rating (n = 56)	0·589	0·334	0·38	0·20–0·57
1 + 2 rating (n = 112)	0·652	0·334	0·48	0·35–0·61
EKS:				
1 rating (n = 56)	0·643	0·302	0·49	0·34–0·64
2 rating (n = 56)	0·518	0·262	0·35	0·20–0·49
1 + 2 rating (n = 112)	0·580	0·280	0·42	0·31–0·52

Table 4 Fifty six adenomas graded according to KMS by two observers on two different occasions (n = 112)

	Mild		*Moderate*		*Severe*	
	Association	*Kappa*	*Association*	*Kappa*	*Association*	*Kappa*
Mild	0·633	0·43	0·367	0·06	0·000	−0·47
Moderate	0·397	0·07	0·466	0·21	0·137	−0·27
Severe	0·000	−0·54	0·139	−0·28	0·861	0·80

Results

INTRAOBSERVER REPRODUCIBILITY

Results of the comparison between the first and second observations by each pathologist are given in table 1 for the KMS and in table 2 for the EKS. The κ values were excellent for the KMS and moderate for the EKS, but this difference was not significant ($p > 0{\cdot}05$).

INTEROBSERVER REPRODUCIBILITY

The overall agreement and reproducibility are given in table 3. Because the cases were coded and the primary diagnoses unknown to the observers, we thought it acceptable to treat the 112 answers (fifty six adenomas graded by two observers on two separate occasions) according to each grading system and as belonging to 112 individual cases from each observer. In our opinion this procedure enabled us to obtain the best estimation of the overall interobserver agreement, which was found to be 0·65 for the KMS and 0·58 for the EKS. The corresponding κ values and their 95% confidence limits were 0·48 (0·35–0·61, KMS) and 0·42 (0·31–

Table 5 Fifty six adenomas graded according to EKS by two observers

	Mild		Moderate		Severe		CIS		IMC	
	Association	*Kappa*	*Association*	*Kappa*	*Association*	*Kappa*	*Association*	*Kappa*	*Association*	*Kappa*
Mild	0·522	0·47	0·478	0·08	0·000	−0·21	0·000	−0·27	0·000	−0·08
Moderate	0·113	0·01	0·722	0·51	0·144	−0·04	0·010	−0·26	0·010	−0·07
Severe	0·000	−0·11	0·359	−0·13	0·410	0·29	0·205	−0·01	0·026	−0·05
CIS	0·000	−0·11	0·021	−0·73	0·167	−0·01	0·583	0·47	0·229	0·17
IMC	0·000	−0·11	0·059	−0·66	0·059	−0·14	0·647	0·55	0·235	0·17

Table 6 Probability for choice among five different categories in EKS when random observer has chosen category in KMS

Chosen category in KMS	*Probability for choice of category in EKS*				
	Mild	*Moderate*	*Severe*	*CIS*	*IMC*
Mild	0·272	0·696	0·032	0·000	0·000
Moderate	0·021	0·514	0·390	0·068	0·007
Severe	0·000	0·063	0·111	0·597	0·229

0·52, EKS). Both κ values differed significantly from $\kappa = 0\cdot75$ ($p < 0\cdot0001$), but not from $\kappa = 0\cdot50$. The association with and reproducibility on the different categories are given for the KMS in table 4 and for the EKS in table 5. The κ values for the KMS showed poor reproducibility for mild and moderate dysplasia, while the reproducibility was excellent with regard to severe dysplasia ($\kappa = 0\cdot80$, SE (κ) = 0·094).

By modifying the KMS into two categories (mild/moderate *v* severe dysplasia), an excellent overall agreement (0·91) and reproducibility ($\kappa = 0\cdot80$, SE (κ) = 0·094) was found. Furthermore, the reproducibility of both categories of this simplified KMS was $\kappa = 0\cdot80$.

For the EKS, the κ values showed poor reproducibility for all categories, except for moderate dysplasia, where the result was moderate. If the two categories of CIS and IMC of the EKS were lumped together, however, the reproducibility of this combined category was excellent ($\kappa = 0\cdot76$, SE (κ) = 0·094. By further simplifying the EKS into two categories (Kozuka's mild/moderate/severe *v* CIS/IMC) the reproducibility was $\kappa = 0\cdot76$.

ASSOCIATION BETWEEN SYSTEMS

The associations between the two systems are given in tables 6 and 7 and show the probability for the choice of a particular category in one system, when the observer has chosen a category in the other system. The associations seem to permit a translation between the two systems as shown in table 8.

Table 7 Probability for choice among three different categories in KMS when random observer has chosen category in EKS

Chosen category in EKS	*Probability for choice of category in KMS*		
	Mild	*Moderate*	*Severe*
Mild	0·935	0·065	0·000
Moderate	0·567	0·387	0·046
Severe	0·064	0·731	0·205
CIS	0·000	0·104	0·896
IMC	0·000	0·029	0·971

Table 8 Translation between EKS and KMS

From EKS	*To KMS*
Mild	Mild
Moderate	Mild/moderate
Severe	Moderate
CIS	Severe
IMC	Severe
From KMS	*To EKS*
Mild	Mild/moderate
Moderate	Moderate/severe
Severe	CIS/IMC

Discussion

The value of a grading system depends on its biological importance as well as its reproducibility. In the past few years a vast number of reports have been published, which indicate that other methods may replace conventional light microscopy for detecting patients at risk. These include morphometry, DNA flow cytometry, immunohistochemistry, lectin binding studies, enzyme histochemistry, cell kinetic studies and others. Although many of these studies have shown promising results in terms of a better understanding of the adenoma-carcinoma sequence, the results have often been inconsistent and until now none has been able to replace conventional histological grading in routine diagnostic pathology.

A test of reproducibility should fulfil several criteria. The number of specimens examined should be sufficiently large[7] and all categories should be represented in a sufficient number to increase the possibility of disagreement.[9] The investigation should be carried out blind and all sections from a given specimen should be included to imitate the diagnostic routine. Furthermore, only with well defined criteria for each category can an acceptable outcome be expected.

When determining what is acceptable, most authors have used the arbitrary division suggested by Landis and Kock,[11] according to which a value of $\kappa \leq 0\cdot20$ should be taken as slight agreement, 0·21–0·40 as fair, 0·41–0·60 as moderate, 0·61–0·80 as substantial and 0·81–1·00 as almost perfect. As the reproducibility of a given variable, by definition, influences the possible level of biological importance of this variable, too low a reproducibility can not be accepted. Therefore we followed the recommendation of Svanholm *et al*,[9] who advocated that $\kappa < 0\cdot50$ should be interpreted as poor reproducibility and $\kappa \geq 0\cdot75$ as excellent reproducibility. Of course, the level of acceptable reproducibility has to be increased the more fundamental the clinical consequences are of a given variable.

Table 9 Reports on κ statistics for assessment of intra- and interobserver reproducibility in estimating colorectal specimens

Histological problem	*No of categories*	*No of investigators*	*Intraobserver kappa*	*Paired interobserver kappa*	*Reference*
Grading of adenocarcinoma	3	2	0·539–0·630	0·115–0·532	Thomas *et al*[12]
Lymphocytic infiltration	2	6	−0·03 –0·52	0·29 –0·30	Dundas *et al*[14]
Invasive margin	2	6	−0·03 –0·82	0·32 –0·66	Dundas *et al*[14]
Dysplasia in IBD	4	6		0·292–0·584	Dixon *et al*[17]
Dysplasia in adenoma	3	2	0·330–0·450	0·230–0·369	Brown *et al*[18]
Dysplasia in adenoma:					
according to EKS	5	2	0·68 –0·70	0·42	Present study
according to KMS	3	2	0·78 –0·81	0·48	Present study

IBD = inflammatory bowel disease

Studies on the reproducibility of estimating histological changes in colorectal pathology have been carried out for several variables in adenocarcinomas and for epithelial changes in inflammatory bowel disease and adenomas. A list of the studies using κ statistics is given in table 9.

Grading of colorectal adenocarcinoma was tested by Thomas *et al*[12] on biopsy specimens, as well as on primary tumours. Intraobserver agreement on three grades was found by two investigators—74 and 80%, with corresponding κ values of between 0·539 and 0·630. Paired interobserver reproducibility among five pathologists showed κ values ranging from 0·115 to 0·532. Intraobserver reproducibility between the findings of biopsy specimens and the corresponding primary tumour showed κ values from 0·249 to 0·420. The latter two results were probably influenced by the fact that no standardised criteria had been accepted by the five observers.

Recently, a new prognostic classification of rectal cancer has been proposed by Jass *et al.*[13] Two of the more subjective variables in this classification were tested for observer variability among six pathologists by Dundas *et al.*[14] While assessing peritumoral lymphocytic infiltration they found an intraobserver reproducibility of κ = −0·03–0·52 and an interobserver reproducibility of κ = 0·29–0·30. Assessment of the invasive margin showed an intraobserver reproducibility of κ = −0·03–0·82 (five pathologists showing κ = 0·44–0·82) and an interobserver reproducibility of κ = 0·32–0·66. The authors concluded that only the latter variable could be reliably assessed.

For inflammatory bowel disease an international group has proposed a "standardised classification" for dysplasia.[15] This classification operates with four grades of dysplasia—namely, probably negative, probably positive, low grade and high grade dysplasia. The authors report good agreement, but still advised that a diagnosis of high grade dysplasia should be reviewed and confirmed before colectomy is performed. The system was tested by Dundas *et al*,[16] who found only minor interobserver and intraobserver disagreements. None of these two studies, however, took chance agreement into account.

In a recent study by Dixon *et al*[17] a system using four categories of inflammatory and dysplastic changes was tested by six pathologists, using κ statistics. Pairwise agreement was found to be from 49% to 72%, with κ values ranging from 0·292 to 0·584. Simplification of the system into two categories, dysplasia *v* no dysplasia, improved the results, but these authors also recommended that a consensus among different pathologists should be obtained before colectomy.

Grading of dysplasia in inflammatory bowel disease differs from grading in adenomas: the biopsy specimens are taken rather at random and the epithelium is under the influence of an eventual inflammation. Dysplasia may be present in a single crypt alone and categories of dysplasia are difficult to define. In contrast, adenomas represent the entire lesion, clinically important inflammatory changes in the epithelium are rare, and categories of dysplasia are well defined. It could therefore be expected that reproducibility would be better.

Dysplasia in colorectal adenomas was studied by Brown *et al*[18] using a three grade system. Two observers obtained intraobserver agreements of 67% to 70%, with corresponding κ values of 0·330 and 0·450, respectively, and their interobserver agreement was 59% to 66%, with κ values of 0·230 to 0·369. The two observers had familiarised themselves with three systems for grading dysplasia,[1-3] but seem not to have chosen among these different systems.

The lower κ values, compared with those in the present study, reflect the importance of choosing one well defined system. Our results clearly show that the terms mild, moderate, and severe dysplasia do not correspond to the same segments of the dysplastic spectrum in the KMS and the EKS.

Grading of adenomas is an artifical division of this spectrum and adenomas often contain areas showing different grades of dysplasia. Using the EKS, an area fulfilling the criteria for a given grade of dysplasia may be very small—that is, a few cells showing complete pseudostratification while in the KMS the grade of dysplasia is based on several criteria, including the architecture of glands and will therefore tend to be based on a larger area. This may imply that such areas will be found more easily at a lower magnification and thereby contribute to explaining why the KMS shows a higher reproducibility. The same phenomenon may in part explain why the diagnosis of IMC was not reproducible in this study. Such areas are very small and only meticulous analysis will indicate loose epithelial cells in the lamina propria.

By simplifying both systems into two groups a high reproducibility can be obtained, but this implies for the EKS that all the original grades (III–V) must be grouped together.

Measurement of reproducibility is an important tool in the selection of classification and grading systems. The potential biological importance of such systems can only be established when the reproducibility is sufficiently good. The results are often disappointing, but we all have a duty to inform our colleagues on the reliability of our diagnoses.

This study underlines the discrepancies between the two grading systems, none of which has impressive overall interobserver reproducibility. By modifying the KMS into two categories (mild/moderate *v* severe dysplasia), however, an excellent overall interobserver reproducibility was found. We therefore propose that this simplified KMS is used not only in diagnostic routine but also when comparing histological changes with other markers for colorectal neoplasia.

This system is only a slight modification of the Consensus from the Working Party on Adenomas, as established in Rome 1988 (Kronborg O, Morson BC, personal communication) and is used in the screening programme for colorectal cancer in Funen, Denmark.[19]

1 Konishi F, Morson BC. Pathology of colorectal adenomas: a colonoscopic survey. *J Clin Pathol* 1982;**35**:830–41.

2 Kozuka S. Premalignancy of the mucosa polyp in the large intestine: I. histologic gradation of the polyp on the basis of epithelial pseudostratification and glandular branching. *Dis Colon Rectum* 1975;**18**:483–93.

3 Ekelund G, Lindström C. Histopathological analysis of benign polyps in patients with carcinoma of the colon and rectum. *Gut* 1974;**15**:654–63.

4 Morson BC, Sobin LH. Histological typing of intestinal tumours. *International histological classification of tumours, vol 15*. Geneva: WHO, 1976.

5 Greaves P, Filipe MI, Abbas S, Ormerod MG. Sialomucins and carcinoembryonic antigen in the evolution of colorectal cancer. *Histopathology* 1984;**8**:825–34.

6 Fenoglio CM, Kaye GI, Lane N. Distribution of human colonic lymphatics in normal, hyperplastic, and adenomatous tissue. Its relationship to metastasis from small carcinomas in pedunculated adenomas with two case reports. *Gastroenterology* 1973;**64**:51–66.

7 Silcocks PBS. Measuring repeatability and validity of histological diagnosis—a brief review with some practical examples. *J Clin Pathol* 1983;**36**:1269–75.

8 Cohen J. A coefficient of agreement for nominal scales. *Education Psychol Measure* 1960;**20**:37–46.

9 Svanholm H, Starklint H, Gundersen HJ, Fabricius J, Barlebo H, Olsen S. Reproducibility of histomorphologic diagnoses with special reference to the kappa statistic. *APMIS* 1989;**97**:688–98.

10 Holman CDJ, James IR, Heenan PJ, *et al*. An improved method of analysis of observer variation between pathologists. *Histopathology* 1982;**6**:581–9.

11 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**: 159–74.

12 Thomas GDH, Dixon MF, Smeeton NC, Williams NS. Observer variation in the histological grading of rectal carcinoma. *J Clin Pathol* 1983;**36**:385–91.

13 Jass JR, Love SB, Northover JMA. A new prognostic classification of rectal cancer. *Lancet* 1987;i:1303–6.

14 Dundas SAC, Laing RW, O'Cathain A, Seddon I, Slater DN, Stephenson TJ, Underwood JCE. Feasibility of new prognostic classification for rectal cancer. *J Clin Pathol* 1988;**41**:1273–6.

15 Riddell RH, Goldman H, Ransohoff DF, *et al*. Dysplasia in inflammatory bowel disease: standardised classification with provisional clinical applications. *Hum Pathol* 1983;**14**:931–68.

16 Dundas SAC, Kay R, Beck S, *et al*. Can histopathologists reliably assess dysplasia in chronic inflammatory bowel disease? *J Clin Pathol* 1987;**40**:1282–6.

17 Dixon MF, Brown LJR, Gilmour HM, *et al*. Observer variation in the assessment of dysplasia in ulcerative colitis. *Histopathology* 1988;**13**:385–97.

18 Brown LJR, Smeeton NC, Dixon MF. Assessment of dysplasia in colorectal adenomas: an observer variation and morphometric study. *J Clin Pathol* 1985;**38**:174–9.

19 Kronborg O, Fenger C, Olsen J, Pedersen KM, Søndergaard O. Preliminary report on a randomised trial of screening for colorectal cancer with Hemoccult II. In: Chamberlain J, Miller AB, *UICC: Screening for gastrointestinal cancer*. Hans Huber Publishers, 1988:41–4.