Modern Pathology (2006) 19, 1236–1242

www.modernpathology.org

Scoring of p53, VEGF, Bcl-2 and APAF-1 immunohistochemistry and interobserver reliability in colorectal cancer

Inti Zlobec[1], Russell Steele[2], René P Michel[3], Carolyn C Compton[4], Alessandro Lugli[3] and Jeremy R Jass[3]

[1]*Department of Pathology, McGill University, Montreal, QC, Canada;* [2]*Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada;* [3]*Department of Pathology, McGill University Health Centre, Montreal, QC, Canada and* [4]*National Cancer Institute, Office of Biorepositories and Biospecimen Research, Bethesda, MD, USA*

Molecular tumor markers are often studied in colorectal cancer using immunohistochemistry to determine their prognostic or predictive value. Protein expression is typically assigned a 'positive' score based on a predetermined cutoff. A semiquantitative scoring method that evaluates the percentage of positive tumor cells (0–100%) may provide a better understanding of the prognostic or predictive significance of these markers. The aim of this study was to assess and compare the interobserver agreement of immunohistochemistry scores using a percentage scoring method and three categorical scoring systems. Immunohistochemistry for p53, Bcl-2, vascular endothelial growth factor (VEGF) and apoptotic protease activating factor-1 (APAF-1) was performed on 87 tumor biopsies from patients with rectal carcinoma and scored independently by four pathologists as the percentage of positive tumor cells. Interobserver agreement was assessed by the intraclass correlation coefficient. The intraclass correlation coefficients for p53 and VEGF (>0.6) indicate substantial agreement between observers. The distribution of Bcl-2 and APAF-1 scores in addition to weaker interobserver agreement by percentage scoring suggest that this approach may not be appropriate for these proteins. In conclusion, p53 and VEGF protein expression assessed by immunohistochemistry in colorectal cancer and scored as a percentage of positive tumor cells may be a viable alternative scoring method.
Modern Pathology (2006) **19,** 1236–1242. doi:10.1038/modpathol.3800642; published online 2 June 2006

Keywords: interobserver reliability; rectal cancer; immunohistochemistry; scoring system; p53; VEGF

Although the TNM stage remains the most significant independent prognostic indicator in patients with colorectal cancer, pathologically identical tumors may neither respond to treatment uniformly nor result in similar survival rates.[1] A number of molecular markers involved in proliferation (p53), apoptosis (Bcl-2, APAF-1) and angiogenesis vascular endothelial growth factor (VEGF) are currently being investigated to determine their value as prognostic or predictive factors and in turn their potential for integration into clinical practice.[2–5]

Immunohistochemistry is an indispensable research and diagnostic tool used to assess the presence or absence of molecular tumor markers on paraffin-embedded tissue.[6] Tumor positivity for a given marker is frequently evaluated using predetermined cutoffs such as 10% ($\leq$10% tumor cells staining = negative, $>$10% = positive).[4,7–10] The employment of categorical scoring systems is motivated by the ease of interpretation of positive tissue by pathologists and is further supported by substantial interobserver agreement. However, it assumes that more detailed analysis of protein expression between 10 and 100%, for example will not contribute any additional relevant information in predicting outcome.[11]

A semiquantitative scoring method that assigns immunohistochemistry scores as a percentage of positive tumor cells (the number of positive tumor cells over the total number of tumor cells) may provide a more complete assessment of protein expression and a clearer understanding of the roles played by potential tumor markers in predicting outcome. Most importantly, by evaluating immunohistochemistry expression semiquantitatively at the

Correspondence: I Zlobec, Department of Pathology, McGill University, 3775 University Street, Room B22, Montreal, QC, Canada H3A 2B4.
E-mail: inti.zlobec@mail.mcgill.ca
Received 31 January 2006; revised 3 May 2006; accepted 4 May 2006; published online 2 June 2006

outset, more relevant cutoffs for tumor positivity may be established for the protein and outcome of interest.

The greatest concern facing such a percentage scoring method is the reproducibility of the scores. In this study, we assess the interobserver agreement of immunohistochemistry scores for four tumor markers known to play a role in progression of colorectal carcinoma and response to radiotherapy namely p53, VEGF, Bcl-2 and APAF-1 and compare the interobserver agreement of percentage scoring to that of three categorical scoring systems.

Materials and methods

In total, 87 pretreatment formalin-fixed paraffin-embedded diagnostic rectal biopsy tissues were collected from a series of patients with rectal adenocarcinoma undergoing preoperative endorectal brachytherapy. [12] Serial sections were cut at 3 μm and immunohistochemistry by the avidin–biotin complex (ABC) procedure, including heat-induced epitope retrieval, was undertaken. Incubation with the primary antibody was carried out in a moist chamber for 1 h at 37°C for p53 (DAKO, clone DO-7, Denmark, 1:100) and at room temperature for VEGF (Santa Cruz Biotechnology, VEGF-A20, USA, 1:100) and APAF-1 (Novocastra, NCL-APAF-1, 1:40). Overnight incubation at 4°C was performed for Bcl-2 (DAKO, clone 124, Denmark, 1:100). Negative controls were treated identically with the primary antibodies omitted. Positive controls consisted of tissue known to contain the protein of interest.

Nuclear positivity for p53 and cytoplasmic positivity for VEGF, Bcl-2 and APAF-1 were evaluated only in areas of invasive carcinoma. Immunoreactivity was scored as the number of positive tumor cells over total tumor cells, independently by four pathologists (CCC, JRJ, RPM, AL); in general each slide took on average 30 s or less to score. No specific instructions or illustrations were presented to pathologists to assist in their evaluation. Percentage scores were subsequently categorized using the 0% cutoff (0% staining *vs* any staining), the 10% cutoff ($\leq$10% tumor cell staining *vs* $>$10% staining) and a three-category scoring system consisting of 0% staining, between 1 and 50% staining and $>$50% staining.

Statistical Analysis

The interobserver agreement for the 0, 10 and 0, 1–50, $>$50% cutoff scoring systems were evaluated using Light's Kappa coefficient.[13] The Kappa coefficient (k) is a useful measure of agreement for categorical data as it takes into account the probability that observers achieved the same scores by chance. General guidelines for the interpretation of Kappa suggest that values between 0.81 and 1.0 should represent 'almost perfect' agreement, 0.61–0.80 'substantial' agreement, 0.41–0.60 'moderate' agreement, 0.21–0.40 'fair' agreement, and 0–0.20 'slight' agreement.[14]

The intraclass correlation coefficient is the most commonly used method to assess interobserver agreement for quantitative measurements.[15] Similar to the simple Pearson correlation coefficient that measures association, the intraclass correlation coefficient additionally estimates agreement between scores from different observers on the same patients. The closer the intraclass correlation coefficient is to 1, the better the agreement between observers. The intraclass correlation coefficient was employed to assess interobserver agreement of percentage scores.

Although no recommendations for the interpretation of the intraclass correlation coefficient have been detailed, reports in the literature have supported the use of the following guidelines: a coefficient of reliability $>$0.75 indicates 'strong' agreement, between 0.4 and 0.75, 'good' agreement, and $<$0.4, 'poor' agreement.[16] It has also been suggested that the values for the Kappa coefficients may be equivalent to the intraclass correlation coefficient making their direct comparison appropriate.[17]

Confidence intervals (95%) were found by 10 000 bootstrap replications of the dataset. All analyses were carried out using SAS Version 8.2 (The SAS System, NC, USA).

Results

p53

Overall mean p53 protein expression was 37% (Table 1). Approximately 72% of tumors were positive for the protein. The frequency distribution of p53 scores was nearly uniform above 0% (Figure 1). The reproducibility of p53 scores was

Table 1 Mean and standard deviation of scores (%) for pathologists 1–4 and overall mean protein expression

	Overall	*1*	*2*	*3*	*4*
p53	36.90±34.09	34.07±33.90	34.43±29.61	32.36±28.67	46.71±41.27
VEGF	45.15±37.69	51.96±39.07	39.26±34.43	31.11±11.03	58.58±39.93
Bcl-2	9.47±22.98	14.16±28.02	9.27±22.33	4.14±13.46	10.06±24.48
APAF-1	17.70±32.21	29.22±39.27	14.85±26.21	2.6±7.99	23.97±38.36

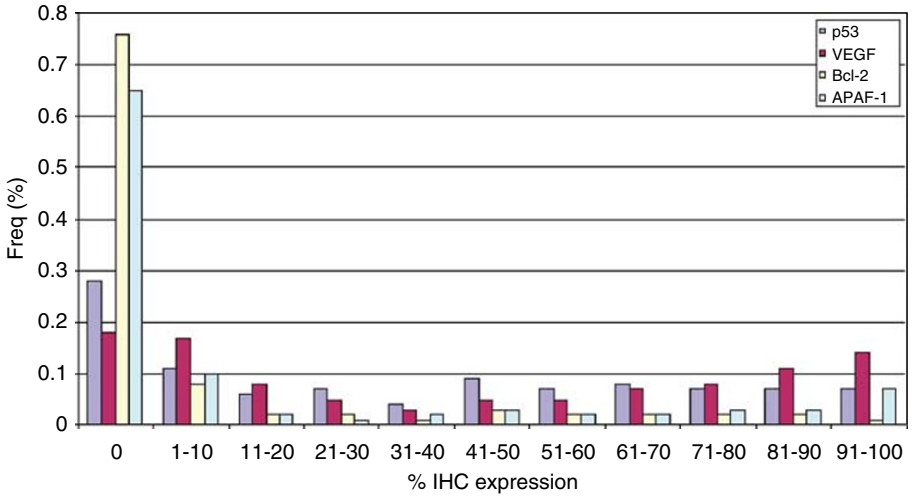


Figure 1 Distribution of p53, VEGF, Bcl-2 and APAF-1 scores.

Table 2 Intraclass correlation coefficient measuring agreement between percentage scores and Kappa coefficients (*k*) measuring agreement of scores using the 0% cutoff, 10% cutoff and 0, 1–50, >50% cutoffs. Intervals represent 95% confidence intervals

	N	*Intraclass correlation coefficient*	*k (0% cutoff)*	*k (10% cutoff)*	*k (0, 1–50, >50% cutoffs)*
p53	86	0.755 (0.67, 0.82)	0.831 (0.73, 0.92)	0.740 (0.63, 0.84)	0.588 (0.48, 0.68)
VEGF	87	0.624 (0.52, 0.71)	0.565 (0.39, 0.71)	0.569 (0.45, 0.68)	0.434 (0.33, 0.53)
Bcl-2	79	0.533 (0.34, 0.69)	0.561 (0.43, 0.68)	0.490 (0.33, 0.63)	0.407 (0.26, 0.55)
APAF-1	85	0.497 (0.41, 0.58)	0.514 (0.40, 0.62)	0.434 (0.33, 0.53)	0.377 (0.30, 0.45)

substantial for both percentage scoring and the 10% cutoff (intraclass correlation coefficient = 0.755 and $k = 0.740$, respectively) (Table 2). Excellent agreement was achieved when no positivity (0%) *vs* any positivity was evaluated ($k = 0.831$). The 0, 1–50, >50% scoring method produced the least amount of agreement between observers. p53 staining was evaluated with less difficulty when no nuclei or nearly all nuclei were positive for the protein (Figure 2a). Staining intensity was generally moderate to strong. Positivity was confined to tumor cell nuclei in the majority of cases. Both the presence of cytoplasmic positivity (Figure 2b) and weak staining intensity in nuclei were largely responsible for the variation in scores.

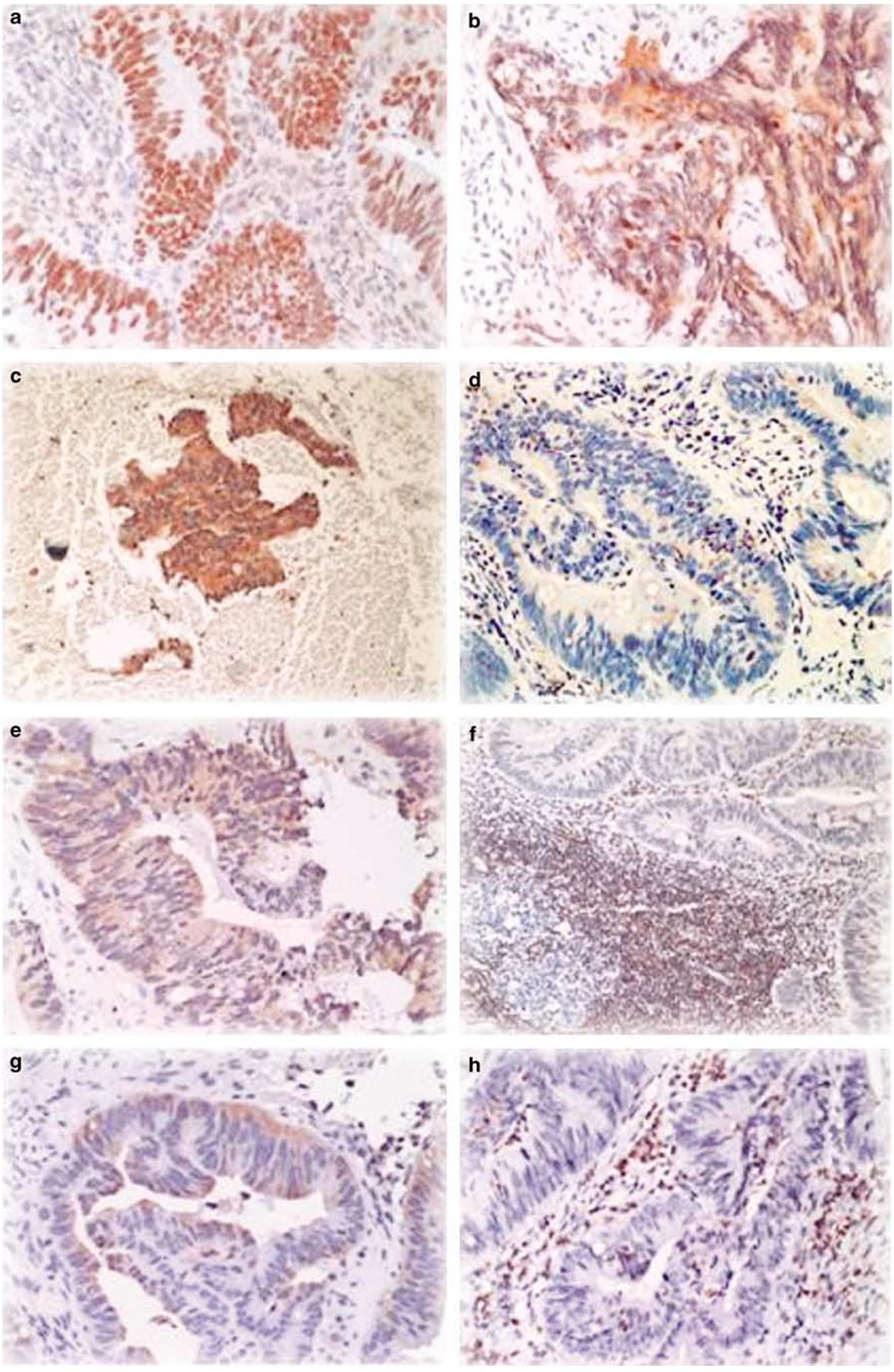
VEGF

The distribution of VEGF scores was U-shaped (Figure 1) with an overall mean cytoplasmic expression of 45% (Table 1). The intraclass correlation coefficient for percentage scoring was 0.624 reflecting a substantial degree of interobserver agreement (Table 2). The categorical scoring systems yielded moderate agreement between observers, the least reproducible being the 0, 1–50, >50% method. The intensity of staining for VEGF varied from weak to strong (Figure 2c). Considerable disagreement between scores could be attributed to weakly stained tumor cells. Infiltration of tumors with a large number of neutrophils may have contributed to the overestimation of the number of positive tumor cells (Figure 2d).

Bcl-2

Approximately 76% of tumors demonstrated complete absence of Bcl-2 (Figure 1). Mean Bcl-2 expression was less than 10% (Table 1). Moderate interobserver agreement was found for percentage scoring as well as for the 0 and 10% cutoffs (Table 2). Agreement was weakest for the 0, 1–50, >50% scoring method ($k = 0.407$). Staining intensity was the primary cause of disagreement of scores between pathologists. Although lymphocytes reacted strongly with the Bcl-2 antibody, only weak to

Figure 2 p53 (**a**, **b**), VEGF (**c**, **d**), Bcl-2 (**e**, **f**) and APAF-1 (**g**, **h**) staining. Tumors in (**a**, **c**, **e** and **g**) resulted in a high degree of interobserver agreement whereas those in (**b**, **d**, **f** and **h**) lead to low interobserver agreement.


a
b
c
d
e
f
g
h

moderate staining was found in tumors expressing the protein (Figure 2e). Infiltration of tumors with large numbers of lymphocytes may have also contributed to disagreement in percentage scores (Figure 2f).

APAF-1

Mean APAF-1 expression determined by each of the four pathologists varied significantly from 2.6 to 29% (Table 1). Approximately 64% of tumors were completely negative for the protein (Figure 1). Moderate agreement was achieved for percentage scoring, as well as for the 0 and 10% cutoffs. The strongest agreement was produced when no staining (0%) *vs* any positive staining was evaluated ($k = 0.514$). APAF-1 positivity was strong in neutrophils and normal mucosa but only weak to moderate staining occurred in tumors expressing the protein (Figure 2g). Substantial neutrophilic infiltration in tumors may have led to disagreement between observers (Figure 2h).

Discussion

The usefulness of any immunohistochemistry scoring method is limited not only to its ability to optimize the prognostic or predictive value of tumor markers but also to its reproducibility. Studies on interobserver agreement in colorectal carcinoma are uncommon. Several studies using the 10% cutoff scoring method describe a high degree of concordance between pathologists evaluating positive and negative tumors.[18–20] This type of agreement typically overestimates true categorical agreement by ignoring the probability that scores were obtained by chance, an important consideration when scores are not evenly distributed as was seen for Bcl-2 and APAF-1 in this study.[21]

The reproducibility of p53 scores either as percentages or by way of the 10% cutoff scoring method was high. Although agreement was strongest at the 0% cutoff, the distribution of p53 expression suggests that it may be important to evaluate the complete range of scores.

The interobserver agreement of percentage scores for VEGF in this study was higher than those for the 0 and 10% cutoffs. The distribution of VEGF scores indicates that percentage scoring may provide additional information about the protein that would otherwise go unrecognized by categorizing positivity according to predetermined cutoffs. We recently demonstrated in patients with rectal cancer undergoing preoperative radiotherapy that mean VEGF expression was significantly higher (63%) in biopsies from patients with nonresponsive tumors than from tumors with complete pathologic response (37%) (P-value $= 0.0035$) hence exemplifying the use of percentage scores.[22]

The reproducibility of Bcl-2 percentage scores was similar to the 10% cutoff. The greatest interobserver agreement was found using the 0% cutoff. Approximately 76% of tumors in this study were completely negative for the protein. This result is in line with the literature which states that the frequency of Bcl-2 expression in rectal carcinoma is less than 30%.[23] Kim *et al*[23] demonstrated that the rate of Bcl-2 overexpression decreases with more advanced Dukes stage. In this study, 98% of rectal biopsies were taken from patients with clinically diagnosed cT3 tumors. This may have biased our results in favor of the 0% cutoff and against percentage scoring as overexpression of Bcl-2 would not be expected to vary significantly in this sample. The interobserver agreement of percentage scores may be better assessed in colorectal adenomas known to frequently overexpress the protein.[23] Our results show that Bcl-2 expression scored as 0% positive tumor cells *vs* any tumor cell staining leads to the highest degree of interobserver agreement in rectal tumors of the same stage.

Recent evidence suggests that *APAF-1* may function as a tumor-suppressor gene.[24] Loss of tumor suppression leads to loss of wild-type APAF-1 protein translating into absence of staining via immunohistochemistry. It is therefore reasonable to suggest that the 0% scoring method with the highest degree of interobserver agreement may be a more meaningful method of evaluation than scoring by percentages for this protein. Although p53 acts as a tumor-suppressor gene as well a similar argument against percentage scoring cannot be used.[25] The short half-life of wild-type p53 renders the protein undetectable to immunohistochemistry.[26] Immunohistochemistry for mutant p53 is based on the assumption that the abnormal protein cannot act as a transcriptional factor hence accumulating in the cell.[25] A comparison or DNA sequencing analysis and immunohistochemistry to detect mutant p53 has revealed a significant false-positive rate for the latter.[25] Immunostaining with p53 antibodies appears therefore to detect abnormal accumulation of p53 in the cell and is not limited to detection of the mutant protein. It is possible that p53 scores evaluated as the percentage of abnormal accumulation of p53 will prove to be a useful predictive factor.

Percentage scoring should allow a more thorough assessment of the predictive or prognostic significance of tumor markers. The correlation between the immunohistochemistry expressions of several proteins can be assessed. Pich *et al*[27] performed percentage scoring of Ki-67, PCNA and MIB-1 expression in non-Hodgkin lymphoma. They found a strong linear correlation for all proteins and used this finding to argue that Ki-67, PCNA and MIB-1 labeling were reliable and complementary methods to assess the proliferative activity of intermediate grade non-Hodgkin lymphoma. By studying the mean expression of Ki-67, PCNA and MIB-1, they identified subtypes of intermediate grade non-

Hodgkin's lymphoma with potentially different prognoses.

Logistic regression is often used to select predictive factors from a pool of possible tumor, host or treatment variables. The risk of development of cancer using serum tumor markers (such as CEA), or the probability of local tumor control with varying doses of radiation are examples of logistic regression with quantitative variables to predict outcome.[28,29] Percentage scoring of immunohistochemistry can be applied similarly to determine how the odds of a binary outcome (response/no response to treatment) change with increases or decreases in protein expression.

Finally, by first quantifying scores, other statistical approaches such as receiver operating characteristic (ROC) analysis can be used to determine the sensitivity and specificity of tumor markers as well as the optimal cutoffs for positivity.[28] By percentage scoring we have shown how classification and regression tree (CART) methods could be used to select proteins playing a role in predicting rectal tumor response to preoperative radiotherapy and to determine the protein cutoff values for optimal discrimination between responsive and nonresponsive tumors.[30]

Percentage scoring of immunohistochemistry expression in colorectal tumors may be suitable for proteins that exhibit a wide range of tumor cell positivity with moderate to strong staining intensity and a high degree of interobserver agreement. The results of this preliminary study on the interobserver agreement of percentage scoring demonstrate that the evaluation of p53 and VEGF using this approach appears to be a reproducible method and viable alternative for the evaluation of immunohistochemistry.

Acknowledgements

We would like to thank Kristi Baker, Professor Sanjo Zlobec and Dr Nilima Nigam for their suggestions and help with editing and Dr Té Vuong for her continued support.

References

1 Compton CC. Colorectal carcinoma: diagnostic, prognostic, and molecular features. Mod Path 2003;16: 376–388.
2 Kahlenberg MS, Sullivan JM, Witmer DD, *et al.* Molecular prognostics in colorectal cancer. Surg Oncol 2003;12:173–186.
3 Russo A, Bazan V, Iacopetta B, *et al.* The TP53 colorectal cancer international collaborative study on the prognostic and predictive significance of p53 mutation: influence of tumor site, type of mutation, and adjuvant treatment. J Clin Oncol 2005;23:7518–7528.
4 Zlobec I, Vuong T, Compton CC. The predictive value of APAF-1 in rectal tumors treated with pre-operative high-dose rate brachytherapy. Cancer 2006;106: 284–286.
5 Boxer GM, Tsiompanou E, Levine T, *et al.* Immunohistochemical expression of vascular endothelial growth factor and microvessel counting as prognostic indicators in node-negative colorectal cancer. Tumour Biol 2005;26:1–8.
6 Taylor CR. An exaltation of experts: concerted efforts in the standardization of immunohistochemistry. Appl Immunohistochem 1994;1:1–11.
7 Rosati G, Chiacchio R, Reggiardo, *et al.* Thymidylate synthase expression, p53, bcl-2, Ki-67 and p27 in colorectal cancer: relationships with tumor recurrence and survival. Tumour Biol 2004;25:258–263.
8 Giatromanolaki A, Stathopoulos GP, Tsiobanou E, *et al.* Combined role of tumor angiogenesis, bcl-2 and p53 expression in the prognosis of patients with colorectal carcinoma. Cancer 1999;86:1421–1430.
9 Kang S-M, Maeda K, Onoda N, *et al.* Combined analysis of p53 and vascular endothelial growth factor expression in colorectal carcinoma for determination of tumor vascularity and lover metastasis. Int J Cancer (Pred Oncol) 1997;74:502–507.
10 Galizia G, Lieto E, Ferraraccio F, *et al.* Determination of molecular marker expression can predict clinical outcome in colon carcinomas. Clin Cancer Res 2004; 10:3490–3499.
11 Cross SS. Grading and scoring in histopathology. Histopathology 1998;33:99–106.
12 Vuong T, Belliveau PJ, Michel RP, *et al.* Conformal preoperative endorectal brachytherapy treatment for locally advanced rectal cancer. Dis Colon Rectum 2002;45:1486–1493.
13 Conger AJ. Integration and generalisation of Kappas for multiple raters. Psychol Bull 1980;88:322–328.
14 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33: 159–174.
15 Shrout PE, Fleiss JL. Intra-class correlations: uses in assessing rater reliability. Psychol Bull 1979;2: 420–428.
16 Coenraads PJ, Van Der Walle H, Thestrup-Redersen K, *et al.* Construction and validation of a photographic guide for assessing severity of chronic hand dermatitis. Br J Dermatol 2005;152:296–301.
17 Fleiss JL, Levin B, Paik MC. The measurement of interrater agreement. In Shewart WA, Wilks SS (eds). Statistical Methods for Rates and Proportions, 3rd edn. Wiley Series in Probability and Statistics Published Online 02 Jan 2004, pp 598–626.
18 Cascinu S, Staccioli MP, Gasparini G, *et al.* Expression of vascular endothelial growth factor can predict event-free survival in stage II colon cancer. Clin Cancer Res 2000;6:2803–2807.
19 Kay EW, Walsh CJ, Whelan D, *et al.* Inter-observer variation of p53 immunohistochemistry—an assessment of a practical problem and comparison with other studies. Br J Biomed Sci 1996;53:101–107.
20 Zu Y, Steinberg SM, Campo E, *et al.* Validation of tissue microarray staining and interpretation in diffuse large B-cell lymphoma. Leuk Lymph 2005;46:693–701.
21 No author. Letter to the Editor: comparison between two test results, *k* statistic instead of simple overall agreement. Vet Parasitol 2005;133:369–370.
22 Zlobec I, Steele R, Compton CC. VEGF as a predictive marker of rectal tumor response to preoperative radiotherapy. Cancer 2005;104:2517–2521.

23 Kim Y-H, Lee J-H, Chun H, *et al.* Apoptosis and its correlation with proliferative activity in rectal cancer. J Surg Oncol 2002;79:236–242.
24 Hajra KM, Liu JR. Apoptosome dysfunction in human cancer. Apoptosis 2004;9:691–704.
25 Munro AJ, Lain S, Lane DP. P53 abnormalities and outcome in colorectal cancer: a systemiatic review. Br J Cancer 2005;1:1–11.
26 Cuddihy AR, Bristow RG. The p53 protein family and radiation sensitivity: Yes or no? Cancer Metast Rev 2004;23:237–257.
27 Pich A, Ponti R, Valente G, *et al.* MIB-1, Ki-67 amd PCNA scores and DNA flow cytometry in intermediate grade malignant lymphoma. J Clin Pathol 1994;47:18–22.
28 Carpelan-Holmstrom M, Louhimo J, Stenman U-H, *et al.* Estimating the probability of cancer with several tumor markers in patients with colorectal disease. Oncology 2004;66:296–302.
29 van Tol-Geerdink JJ, Stalmeier PF, Pasker-de Jong PC, *et al.* Systematic review of the effect of radiation dose on tumor control and morbidity in the treatment of prostate cancer by 3D-CRT. Int J Radiat Oncol Biol Phys 2006;64:534–543.
30 Zlobec I, Steele C, Nigam N, *et al.* A predictive model of rectal tumor response to pre-operative radiotherapy using classification and regression tree methods. Clin Cancer Res 2005;11:5440–5444.