Research in Surgery, Vol. 4, No. 1, April 1992

RESEARCH
IN
SURGERY

Quantification of Variation in Reading Immunohistochemical Assays

F. Michelassi[1], W. Pomidor[3], A.G. Montag[2], J. Stephens[2], R.D. Goldberg[2], M. Handcock[4]
Departments of [1]Surgery and [2]Pathology, University of Chicago Medical Center, Chicago IL.
[3]Department of Clinical Research, St. Elizabeth Hospital Medical Center, Youngstown OH.
[4]Department of Statistics, New York University, NY, USA.
Address: F. Michelassi. University of Chicago, Department of Surgery, 5841 S. Maryland Avenue, Chicago, Illinois 60637. USA.

ABSTRACT

Semiquantitative immunohistochemical assays have been used with increasing frequency. This study was designed to investigate the reproducibility of such measurements by observing how the measured level varied due to (a) the choice of tissue sample from a single or multiple tumors, (b) the immunohistochemical procedure and the influence of time on staining and (c) the subjective variability between readers (interobserver) and by the same reader (intra-observer). The study was meant to judge the reproducibility of the method, not its accuracy. The choice of the monoclonal antibody therefore did not influence the results. A total of 128 sets of slides from 8 colonic adenocarcinomas were analyzed by three pathologists using a randomized, symmetric, prospective, doubleblind study. There was surprisingly poor agreement between readings of the same case by the three pathologists (37%) and by the same pathologist over time (58%). Based on the component of variation analysis, 11% of the total variation was due to differences in the immunohistochemical procedure, 5% to variation of expression in different tumors, 5% to interobserver and 79% to intra-observer variability. Readings of semiquantitative immunohistochemical assays is limited by subjective intrinsic variability.

Key words: Immunohistochemistry, assay. Adenocarcinoma.

INTRODUCTION

Semiquantitative immunohistochemical assays have been used with increasing frequency in laboratory and clinical practice. We have used this methodology extensively in the past for the study of *ras* oncogene expression. Using the RAP-S monoclonal antibody, we have demonstrated that the degree of *ras* oncogene expression parallels the malignant potential of benign colonic lesions (l); also, that colonic and rectal cancers with increased *ras* oncogene expression are more likely to develop distant metastases and that the overall survival of such patients after curative resection is lower (2,3).

In these studies, the level of *ras* oncogene protein product (p21) was measured using RAP-S monoclonal antibody in a semi-quantitative immunohistochemical assay. The level of expression was defined as the highest serial dilution of antibody giving a definite staining with the avidin-biotin peroxidase method. Each sample was then read by at least two investigators with all differences settled by a third one.

However, the reproducibility of *ras* oncogene protein product measurements using the RAP-S monoclonal antibody in a semi-quantitative immunohistochemical assay had never been investigated. This study was

Intra-observer variation	79%
Inter-observer variation	5%
Experimental variation	11%
Tumor variation	5%

Table I. Relative contribution to variation by each factor in the aggregate analysis model.

Intra-observer variation	49%
Inter-observer variation	16%
Experimental variation	18%
Tumor variation	17%

Table II. Relative contribution to variation by each factor in the dichotomous analysis model.

designed to assess variation in measured levels of expression resulting from variation of expression within tissue samples, variation within tumors, the immunohistochemical procedure and finally, differing judgments between individual readers and by the same reader over time. The study was not designed to judge the precision of the method or its accuracy.

Rather than investigating the correlation between titers and levels of *ras* oncogene expression or the specificity of the RAP-S monoclonal antibody, we studied the reproducibility of the semi-quantitative immunohistochemical method and attempted to identify the potential sources of error.

MATERIAL AND METHODS

Eight colonic adenocarcinomas from 8 different patients formed the basis of this prospective, randomized, doubleblind, symmetric study. Three 1 x 1 x 1 cm samples ("blocks") were randomly obtained from each tumor by the technical staff of the Laboratory of Surgical Pathology; a fourth block was collected by our laboratory technician to check for sampling variation. The inclusion of different blocks from the same tumor provided a mechanism for determining whether different areas of a given tumor had different levels of oncogenic expression.

A total of four cases were sampled from each block: two cases and their replicates (Fig. 1). This technique enabled analysis for variability of expression in nearby histological slices within the same block. Also, because the replicates were stained on different days, we were able to check for variability due to the staining procedure itself.

This study design resulted in 128 cases (8 tumors x 4 blocks x 4 cases). Each case was composed of a set of 6 slides which were stained at progressively higher dilutions of RAP-S monoclonal antibody using an avidin-biotin immunohistochemical assay. Details of the method have been thoroughly described in previous publications (1-3).

All 128 cases were read independently by three pathologists in a randomized order. Each pathologist determined the highest dilution of antibody giving a definite staining. The participation of three pathologists provided the means for determining inter-observer variability.

Each pathologist required an average of two weeks to complete a reading of all 128 cases. In order to test for intra-observer variation and the effect of time on the immunohistochemical staining, each pathologist read the complete set of cases a second time in a different randomized order. The data were collected over a 7 month interval.

Within each case, the 6 RAP-S monoclonal antibody dilutions were presented sequentially to the reader. However, the clinical and pathologic characteristics of the tumor being read were unknown to the reader at the time of the reading. Results were recorded as <1:5000, 1:10,000, 1:20,000, 1:40,000, 1:80,000, and 1:160,000. When there was ambiguity about the result, a "+/-", or a written note was recorded beside the result. The initial exploratory univariate analysis and the following multifactorial model were in agreement with the statistical methodology described by McCullagh (4), and McCullagh and Nelder (5).

RESULTS

The project was completed without lost or ruined slides. However, 25 cases were labeled by a reader as "no tumor present, "very little tumor present, “tumor fragmented", "slides difficult to read", "inconsistent slides", or "+/-." Since in practice the tumor would be resampled and the *ras* oncogene expression process repeated, these cases were excluded from the results. Consequently, the total number of readings considered in the analysis was reduced from 768 (128 cases x 3 readers x 2 readings) to 618 readings. The few readings that were markedly different from all other readings were identified. Because these readings presumably represented error which could occur in practice, they were included in the results.

An exploratory data analysis was performed that evaluated only the effects of intra- and inter-reader variability. The results of this analysis indicated that there was poor agreement between readings for the same case and the same reader over time. In only 58% of the cases was the second reading identical to the first reading for the three readers (Fig. 2). In 90% of the cases the second reading was within one dilution of the first reading. There was also low inter-reader consistency. In only 37% of cases did the first reading of all three readers agree. Similarly, in only 36% of cases did the second readings of the three readers agree.

The effect of the time interval between staining and reading the specimen was also assessed. The first reader completed both readings soon after the staining was performed, the second had a 7 month gap between the first and second reading, while the third made both readings 7 months after the staining. Overall and for the same reader, there were no definite effects due to the interval between the staining procedure and the readings. However, any small effect due to the delay between staining and reading could have been obscured by the high overall variability in the readings. Interestingly, there was a trend common to all three readers to call definite staining in lower dilutions during the second reading than during the first. However, when analyzed in a multifactorial model, this

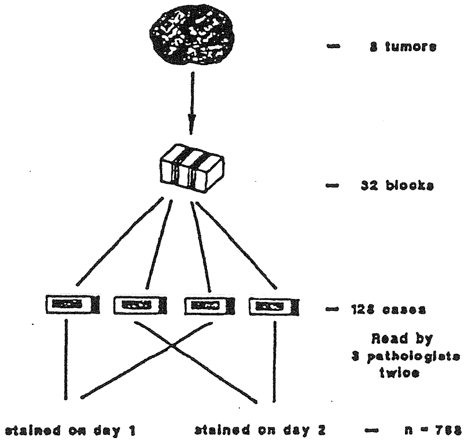

Fig. 1. Scheme of design study.

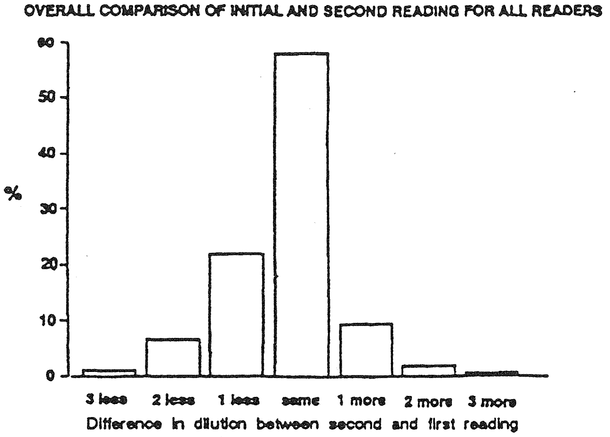

Fig. 2. Overall comparison of first and second reading for all readers.

finding failed to attain statistical significance.

In order to evaluate the contribution of each factor studied on the total variability in the assay, a multifactorial model was developed. Because the study employed random sampling techniques (with the exception of the readers), the factors were regarded as "random effects" drawn from the population of interest. For instance, because the tumors were selected at random, it was assumed that each tumor was from a distribution, with a variance characteristic of the entire population of such tumors. Because the time of the reading was found to have no clear effect on the variation, the effect of time on the staining was represented as a non-random effect.

The results indicated there was little evidence of interactions between the major effects. In addition, the degree of variation was minimal between blocks from the same tumor and between replicates in the same blocks. Because of these findings, readings from replicates and blocks within the same tumor were pooled. A simpler components of variation model was developed that incorporated reader (inter-observer) variability, the immunohistochemical staining procedure and tumor. The remaining variability (intrinsic variability), after these factors were considered, was attributed to the variation due to the reader (intraobserver variability).

The results from this model (Table I) indicated that the variation between readers and between tumors was each only 5%. The variation due to differences in the immunohistochemical staining procedure was 11%. There was no significant effect due to the delay between reading and staining. Therefore, the remaining variation, roughly 80% of the total variation, was attributable to differences of the reader looking at the same case at different times. Because the exact determination of stain level might not be clinically significant, the data were analyzed to see whether the reproducibility of the test would improve with a dichotomous analysis, using 1:40,000 dilution as the cutoff point. Again, replicates and blocks within the same tumors were aggregated. Table 2 shows that the reader, staining procedure and tumor each contribute about one-sixth of the total variation in making the diagnosis. The remaining one-half of variation can be attributed to the variation in the judgement of the degree of staining for the same reader looking at the same slide at a different time.

DISCUSSION

This experiment, which used randomization and "blindness" wherever possible, was designed to investigate the reproducibility of the semiquantitative immunohistochemical measurements of *ras* oncogene expression for adenocarcinoma. It was not designed to evaluate the sensitivity or the specificity of the *ras* oncogene assay. Instead, it was to explore the factors and their relative contributions to the variation observed in the reading of the stained specimens.

The results indicated that the variation in the judgement of the same reader looking at the same slide was the major contributor to the variation in measurements. Using a multifactorial model, reader inconsistency was associated with about 80% of the total variation seen between readings. The contributions for replicates from the same block and from the same tumor were negligible. The immunohistochemical procedure and the differences between readers contributed only 11% and 5% of the variability, respectively, while differences between the tumors themselves contributed only 5%. In particular, the individual tumor studied was only a minor factor in the differences between the readings.

Because much of the research using semiquantitative immunohistochemical measurements involves choosing a specific level of staining to indicate clinical significance, the data were analyzed to see whether the reproducibility of the test would improve with a dichotomous analysis. The 1:40,000 dilution was chosen as a cutoff point, because in prior work with the *ras* oncogene assay, this level was used as a possible prognostic indicator. The results from this analysis were similar: one-half of the total variation remained attributable to intra-reader variation.

Marked intra- and inter-observer variability have been also noted for the diagnosis of neoplasms or dysplasia when the degree of the abnormality present is rated subjectively "by eye". Significant variability has been observed, especially in the diagnosis of indefinite or low-grade changes (6-8). Reid et al. (6), reported that intra-observer agreement was highest (average 88%) in choosing between high-grade dysplasia or intramucosal carcinoma in Barrett's esophagus versus another diagnosis. The intraobserver agreement was lowest (74%) when asked to distinguish between lesions that were more closely related: negative versus indefinite or low grade dysplasia versus high grade dysplasia or intramucosal carcinoma. Ismail et al. (7) described similar findings: intra-observer agreement was excellent for invasive carcinoma of the cervix and high-grade lesions, it was mediocre to poor for normal and low-grade lesions.

Both studies also found that intra-observer agreement was consistently better than inter-observer agreement. However, neither study analyzed the variability with a multifactorial model. Their findings are in agreement with our exploratory findings of 58% intra-observer and 37% inter-observer consistency when analyzed as independent factors.

Other studies confirm the difficulty in subjectively distinguishing between categories that are based on apparently continuous changes (9-11). For example, Rosai (9) looked at breast pathology for which a diagnosis was required based on subjective assessment of surgical pathology slides. He asked 5 experienced highly respected surgical pathologists for a diagnostic opinion on 17 cases of borderline epithelial lesions of the breast. He found marked inter-observer variability: there was not one case in which all 5 pathologists agreed on a diagnosis and only three cases (18%) in which four of the 5 agreed. Similar results were also reported by DeVet (10) and associates who asked four pathologists to grade 106 specimens of dysplasia. They found considerable disagreement among pathologists in distinguishing between adjacent categories of dysplasia for all grades of dysplasia. They concluded that grading is difficult when based on arbitrarily determined distinct categories of a continuous process that lacks natural and sharp borders.

We also found that readers had difficulty judging between adjacent categories. Although in only 58% was there intra-reader concordance between the two readings, in 90% of cases, the second reading was within one dilution of the first reading. Unfortunately, choosing the adjacent category can dramatically influence prediction of prognosis. For example, for rectal and colon cancer (2,3) we dichotomized the *ras* oncogene measurement at a level of 1:40,000 and found that the higher levels correlated significantly with a worse clinical outcome. However, the results from this study indicated that there was significant intraobserver variability also in the dichotomous model.

Therefore, the results of this experiment suggest that measurements based on readings "by eye" of immunohistochemical assays are limited by the marked reader variability, and that alternative methods need to be considered.

REFERENCES

1. Michelassi F, Leuthner, Lubienski M et al. (1987). *Ras* oncogene p21 levels parallel malignant potential of different human colonic benign conditions. *Arch Surg*, 122: 1414-1416.
2. Michelassi F, Grad G, Erroi F et al. (1990). Relationship between *ras* oncogene expression and clinical and pathological features of colonic carcinoma. *Hepatogastroenterology*, 37(5): 439-442.
3. Michelassi F, Vannucci LE, Montag A et al. (1988). *Ras* oncogene expression as a prognostic indicator in rectal adenocarcinoma. *J Surg Res*, 45:15-20.
4. McCullagh P (1980). Regression models for ordinal data. *JR Statist Soc*, 42: 109-42.
5. McCullagh P, Nelder JA (1989). In: Generalized lineal models. Second edition, New York, Chapman and Hall, pp. 72-99.
6. Reid BH, Haggitt RC, Rubin CE et al. (1988). Observer variation in the diagnosis of dysplasia in Barrett's esophagus. Hum Pathol, 19(2): 166-178.
7. Ismail SM, Colclough AB, Dinnen JS et al. (1990). *Histopathology*, 16: 371-376.
8. Fenger C, Bak M, Kronborg O et al. (1990). Observer reproducibility in grading dysplasia in colorectal adenomas: Comparison between two different grading systems. *J Clin Pathol*, 43: 320-324.
9. Rosai J (1991). Borderline epithelial lesions of the breast. *Am J Surg Pathol*, 15(3): 209-211.
10. DeVet HCW, Knipschild PG, Schouten HJA et al. (1990). Interobserver variation in histopathological grading of cervical dysplasia. *J Clin Epidemiol*, 43(12): 1395-1398.
11. Argyle JC, Benjamin DR, Lampkin B et al. (1989). Acute non-lymphocytic leukemias of childhood. *Cancer*, 63: 295-301.