Reproducibility of the Measurement of Colonic Proliferation Using Bromodeoxyuridine across Laboratories[1]

Denise J. Roe,[2] David S. Alberts, Michael J. Wargovich, Roberd M. Bostick,[3] Harinder S. Garewal, Janine Einspahr, Lisa Fosdick, Lois Ramsey, Cynthia Woods, and Daniel L. McGee[4]

Arizona Cancer Center [D. J. R., D. S. A., H. S. G., J. E., D. L. M.] and Departments of Family and Community Medicine [D. J. R.], Medicine and Pharmacology [D. S. A], and Radiation Oncology [D. L. M], College of Medicine, The University of Arizona, Tucson, Arizona 85724; Department of Gastrointestinal Medical Oncology and Digestive Diseases, M. D. Anderson Cancer Center, University of Texas, Houston, Texas 77030 [M. J. W., C. W.]; Divisions of Epidemiology [R. M. B.] and Biostatistics [L. F], School of Public Health, University of Minnesota, Minneapolis, Minnesota 55414; and Department of Veteran's Affairs Medical Center, Tucson, Arizona 85723 [H. S. G, L. R.]

Abstract

Although measures of colonic cell proliferation are being used as potential intermediate markers in chemoprevention studies, measurement standardization is still ongoing. This study was designed to assess the reproducibility of the labeling index quantification, as measured by bromodeoxyuridine, across four laboratories experienced in its use. Each institution submitted 10 slides, with one circled area of each slide to be scored. Each site followed its standard procedures for scoring colonic crypts; no attempts to standardize these procedures were made. There was high concordance among the laboratories on whether scorable crypts were present on a particular slide, but only two pairs of laboratories demonstrated agreement statistically greater than that predicted by chance. The overall difference among the sites on the number of scorable crypts was marginally significant ($P = 0.083$), and there was a highly significant overall difference in the magnitude of the labeling index ($P < 0.0001$). Sites 1 and 2 tended to have similar results, as did sites 3 and 4, most likely due to common training. Even with these discrepancies, high correlation ($r > 0.75$) was observed among the reported labeling index values for each pair of laboratories. Without standardized training, these laboratories may differ in the crypts considered appropriate for counting and in whether cells are counted as labeled or unlabeled. These results suggest that standardized training in scoring across all sites performing labeling index determinations is required to assure reproducibility across sites or studies. These results may also help explain discrepancies in the average values of the labeling index reported in the literature.

Introduction

Measurement of colonic cell proliferation has been used increasingly in studies of chemopreventive agents. These measurements are presumed to reflect changes in the colonic mucosa, which may eventually lead to colorectal cancer. The rapid growth of interest in these measurements is due to their ability to dramatically shorten the time required to test a novel agent, both due to decreases in the required number of subjects and the time required for observation of the relevant end point (1). Coupled with this growth has been the increasing recognition that the behavior of these intermediate markers must be better understood if they are to be used most effectively (2, 3).

One measure of colonic proliferation is the labeling index, as measured by BrdUrd[5] uptake into epithelial cell DNA (4). BrdUrd preferentially labels cells in the S-phase. The labeling index is the ratio of the number of labeled cells divided by the total number of cells in a colonic crypt; regions exhibiting greater proliferation should have higher values of the labeling index.

Several studies have addressed components of the reproducibility of the labeling index (5–8). Physiological components have included the reproducibility of results within the same subject from multiple biopsies at the same time and from repeated biopsies at a 4-week interval (5). Laboratory components have included the reproducibility of results for the same specimen across different readers within the same laboratory (5–7, 9, 10). However, the reproducibility of results across laboratories has not been addressed. This is a key issue in validating the labeling index as an intermediate marker, because substantial differences across laboratories would suggest a lack of standardization in quantifying the results. Such differences could potentially contribute to the discrepant labeling index values observed in chemoprevention studies (7). To assess the reproducibility of scoring to quantify the labeling index, we designed a round-robin study in which experienced technicians from four laboratories scored the same set of colon biopsy slides immunohistochemically processed for BrdUrd.

Materials and Methods

Slide Selection. The following four institutions participated in the study: M. D. Anderson Cancer Center (Michael J.

Received 9/6/95; revised 2/14/96; accepted 2/14/96.
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
[1] This work was supported in part by USPHS Grants CA-41108, CA-23074, and CA-51932 from the National Cancer Institute, NIH, Department of Health and Human Services.
[2] To whom requests for reprints should be addressed, at Arizona Cancer Center, 1515 N. Campbell Avenue, Tucson, AZ 85724.
[3] Present address: Department of Public Health Sciences, Bowman Gray School of Medicine, Wake Forest University, Winston-Salem, NC 27157.
[4] Present address: Department of Preventive Medicine and Epidemiology, Loyola University Medical Center, Maywood, IL 60153.

[5] The abbreviations used are: BrdUrd, bromodeoxyuridine; PCNA, proliferating cell nuclear antigen; PI, principal investigator.

Wargovich, PI), University of Minnesota (Roberd M. Bostick, PI), University of Arizona (David S. Alberts, PI), and the Tucson Veterans Affairs Medical Center (Harinder S. Garewal, PI). Each institution submitted 10 slides of normal-appearing tissue obtained at least 10 cm above the anal verge from patients with histologically confirmed sporadic adenomas or resected colon cancer. Each institution circled one area they would routinely score. Preparation of the slides followed the standard procedures in place at each institution; no special preparation was requested. Generally, biopsies of the rectal mucosa were obtained from study subjects using flexible sigmoidoscopy; they were then routinely processed, paraffin embedded, cut in serial sections, placed on glass slides, and immunostained using anti-BrdUrd antibody. Details of this procedure for each laboratory are given elsewhere (5, 11, 12). The slides were then distributed to each site for review.

Analysis of Crypts. Each site was asked to follow its standard protocol for scoring crypts; no attempts to standardize these procedures were made. Generally, crypts were considered scorable if they were U shaped, extended from the muscularis mucosae to the lumen, and contained a continuous single layer of epithelial cells. Crypts were divided into two crypt columns at the midpoint of the crypt base; each side contained a column of cells from the base to the lumen. Three of the four sites considered crypts the units of observation and only counted full crypts (both crypt columns), whereas one site (site 3) considered crypt columns the units of observation and included counts of both full crypts and crypt columns. The number of crypts (either full crypts or crypt columns) considered adequate for scoring was determined for each slide. For each crypt (or crypt column), the total number of cells and the number of labeled cells were determined. Cells were considered labeled if they demonstrated any staining, irrespective of its faintness. The labeling index for each slide was then computed: sites 1 and 2 computed the labeling index separately for each crypt (number of labeled cells/total number of cells), then averaged them across crypts, whereas sites 3 and 4 computed the labeling index as the total number of labeled cells (from all crypts) divided by the total number of cells (from all crypts). Note that the first method weights each crypt equally, whereas the second method weights each crypt proportionally to the number of counted cells. The difference between these two methods, therefore, will be larger when there is greater variation in the number of cells counted per crypt.

Statistical Analysis. The statistical analysis focused on agreement in the number of scorable crypts and the magnitude of the labeling index. Pairwise agreement among the four sites on the presence of scorable crypts was assessed using a κ statistic (13); when discordance was present, its association with a particular site was tested using McNemar's test (14). The number of scorable crypts was transformed by the logarithm (+1) to generate values that were more normally distributed; the distribution of the observed labeling index was normal; therefore, these values were not transformed. Overall differences among the sites for the transformed number of scorable crypts and the magnitude of the reported labeling index were tested using repeated measures ANOVA (15). If this overall difference was statistically significant, pairwise comparisons were tested using contrasts of the appropriate sites. Because these contrasts were based on the variability estimates obtained using all four sites, they may yield slightly different results from those obtainable using paired t tests. Adjustment for multiple comparisons was performed using the Bonferroni procedure (16); all P values given are unadjusted unless otherwise noted. Plots of the observed differences in the reported labeling index *versus* the observed differences in the number of scorable crypts were generated for each pair of sites, to assess whether there was a detectable pattern in these differences. Overall differences among the sites in the average number of labeled cells per crypt and the average number of total cells per crypt were also tested using repeated measures ANOVA; correlations between differences in these factors and differences in the reported labeling index were estimated using the Pearson correlation coefficient (14). Finally, the correlation of the reported labeling index values between each pair of sites was estimated using the Pearson correlation coefficient, with a 95% confidence interval (14).

Table 1 Agreement on the presence of scorable crypts

Values on the top line are the entries for each two-by-two table in the format a, b, c, and d (a is the number of slides that both sites considered scorable; b is the number of slides that only the row site considered scorable; c is the number of slides that only the column site considered scorable; and d is the number of slides that neither site considered scorable), followed by the estimated κ statistic (unadjusted P for the test that agreement was greater than expected by chance). For example, both sites 1 and 2 considered 31 slides scorable; site 1 only considered 3 slides scorable; site 2 only considered 4 slides scorable; and both sites considered 2 slides nonscorable, resulting in an estimated κ statistic of 0.263 (P = 0.047).

	Site 1	Site 2	Site 3	Site 4
Site 1		31,3,4,2 0.263 (0.047)	33,1,3,3 0.546 (0.0002)[a]	30,4,4,2 0.216 (0.086)
Site 2			34,1,2,3 0.625 (<0.0001)[a]	30,5,4,1 0.053 (0.37)
Site 3				32,4,2,2 0.318 (0.019)
Site 4				

[a] Agreement statistically greater than predicted by chance ($P < 0.05$) after adjustment for multiple comparisons.

Results

Data analysis addressed four issues. First, is there agreement on whether scorable crypts are present on a particular slide? Second, is the number of scorable crypts counted the same? Third, is the magnitude of the labeling index similar? Finally, what is the correlation between the labeling indices reported by each laboratory?

Of the 40 slides considered, all sites agreed that there were no scorable crypts present for 1 slide (2.5%); all sites agreed that scorable crypts were present for 27 slides (67.5%); and at least one site decided there were no scorable crypts present for the remaining 12 slides (30.0%). Table 1 shows the agreement on the presence of scorable crypts for each pair of sites. As can be seen, the number of concordant pairs ranged from 31 to 37 slides. After adjusting for chance agreement, however, the agreement between the pairs of sites was lower, with values of the κ statistic ranging from 0.053 (site 2 *versus* site 4) to 0.625 (site 2 *versus* site 3); only two pairs of sites demonstrated agreement statistically greater than that predicted by chance (site 1 *versus* site 3, adjusted P = 0.0012; site 2 *versus* site 3, adjusted P = 0.0002). When discordance was present, there was no association between the inability to detect scorable crypts and a particular site; the smallest P value from the six pairwise comparisons was 0.63.

Table 2 Number of crypts counted and labeling index for each laboratory

Values on the diagonal are the mean (SD) for each site; values on the off-diagonal are the mean (SD) of the difference between the values reported for each pair of sites. For example, the mean numbers of crypts counted were 2.325 (SD, 1.992) for site 1 and 2.350 (SD, 2.155) for site 2. The mean difference between these two sites was −0.025 (SD, 1.761).

	Site 1	Site 2	Site 3[a]	Site 4
Mean no. of crypts counted/slide and mean difference between sites (SD)				
Site 1	2.325	−0.025	0.700	0.425
(n = 40)	(1.992)	(1.761)	(1.344)	(1.551)
Site 2		2.350	0.725	0.450
(n = 40)		(2.155)	(1.826)	(1.280)
Site 3[a]			1.625	−0.275
(n = 40)			(1.170)	(1.569)
Site 4				1.900
(n = 40)				(1.837)
Mean % labeling index/slide and mean difference between sites (SD)				
Site 1	6.367	−0.187	−2.255	−1.577
(n = 34)[b]	(2.718)	(1.626)	(2.179)	(2.090)
Site 2		6.330	−1.903	−1.142
(n = 35)[b]		(3.099)	(2.327)	(1.840)
Site 3			8.259	0.813
(n = 36)[b]			(4.004)	(2.006)
Site 4				7.934
(n = 34)[b]				(3.422)

[a] Includes single crypt columns counted as a full crypt.
[b] No. of slides with scorable crypts.

Table 2 provides the means and SDs of the number of crypts counted for each laboratory and the differences between each pair of laboratories. Although statistical analysis was based on the logarithm of the observed number (+1), the untransformed values are displayed in Table 2 for descriptive purposes. As can be seen, sites 1 and 2 tended to count larger numbers of crypts than sites 3 and 4. After transformation, the difference in the number of crypts counted among the sites was marginally significant ($P = 0.083$).

The means and SDs of the labeling index for each laboratory and the differences between each pair of laboratories are also shown in Table 2. Again, note the similarity of results in sites 1 and 2 *versus* sites 3 and 4. The difference in the magnitude of the labeling index across all sites was highly significant ($P < 0.0001$). There was no significant difference between sites 1 and 2 (unadjusted $P = 0.39$) or sites 3 and 4 (unadjusted $P = 0.0710$). However, site 3 differed from sites 1 (adjusted $P < 0.0001$) and 2 (adjusted $P < 0.0001$), whereas site 4 also differed from sites 1 (adjusted $P < 0.0001$) and 2 (adjusted $P = 0.0048$).

One possible explanation for the similar results observed for sites 1 and 2 *versus* sites 3 and 4 is commonalties in the method of labeling index computation. Sites 1 and 2 computed the labeling index separately for each crypt, then averaged them across crypts, whereas sites 3 and 4 computed the labeling index as the total number of labeled cells (from all crypts) divided by the total number of cells (from all crypts). Table 3 provides the mean and SD of the percentage labeling index for each site computed using each method, with the largest observed negative and positive differences. As can be seen, the mean difference between the methods was quite small, certainly not large enough to explain the differences observed in Table 2. Additionally, Table 3 suggests that the between-slide variability is virtually identical with the two methods.

An obvious question is whether the observed differences in the reported labeling index between the laboratory pairs were associated with the observed differences in the number of scorable crypts. There was no detectable pattern in these results; similar differences in the labeling index were reported for slides with the same numbers of scorable crypts and for slides with different numbers of scorable crypts. Alternative explanations could be differences in the number of labeled cells or differences in the total number of cells counted per crypt across laboratories. Although there were highly significant differences in both the average number of labeled cells ($P < 0.0001$) and total cells ($P < 0.0001$) per crypt, only differences in the average number of labeled cells showed statistically significant correlations with the differences in the labeling index (all $r > 0.63$). A minimal correlation was observed between the labeling index differences and differences in the average number of total cells per crypt ($-0.38 < r < 0.06$).

Plots of the magnitude of the labeling index reported for each pair of sites are shown in Fig. 1; the plots display the observed values with the line of identity, the estimated correlation coefficient, and its 95% confidence interval. As can be seen, the correlations among the sites are high (all $r > 0.75$). However, differences in the magnitude of the values reported by sites 3 and 4 can be appreciated by observing the larger number of observations that lie below the line of identity.

Discussion

Several studies have explored the factors that influence variability in proliferative measures (2, 5). These factors include differences in the genetic and dietary backgrounds of study subjects, differences in disease states, differences in preparation for study procedures, differences in handling and processing study specimens, and differences in quantifying the results. Of these factors, adjusting for complex between-subject differences may be the hardest to potentially modulate, whereas factors related to study procedures may be easier. Therefore, we chose to focus this study on the reproducibility of scoring the BrdUrd-labeling index across multiple laboratories.

To summarize our findings, although there was high overall concordance among the laboratories on whether scorable crypts were present on a particular slide, only two pairs of laboratories demonstrated agreement statistically greater than that predicted by chance. Additionally, the difference in the number of scorable crypts was marginally significant, and there was a highly significant difference in the magnitude of the reported labeling index. Generally, sites 1 and 2 tended to be more similar, as did sites 3 and 4. Differences in the reported labeling index were strongly correlated with differences in the average number of labeled cells per crypt but were not correlated with differences in the number of crypts or differences in the average number of total cells per crypt. However, even with these discrepancies, a high correlation was observed among the reported labeling index values. Thus, although the reported magnitude differed among the sites, there was agreement in the ranking of the labeling index values.

The most plausible explanation for this pattern is common training. On review of these results, we discovered that the site 2 technician had been trained by the laboratory personnel of site 1 and that the site 4 technician had been trained by the laboratory personnel of site 3. This can easily explain the greater concordance among these pairs of sites, especially in the number of scorable crypts and the magnitude of the reported labeling index.

Although we explored the alternative explanation that this pattern was due to commonalties in the method of the labeling index computation, the mean difference between the methods was quite small. Results from at least two studies suggest that

Table 3 Comparison of labeling index computational methods

Values shown are the mean (SD) for each site using each computational method, the mean (SD) of the difference between the values reported using each computational method, and the largest observed negative and positive difference.

	Average across crypts	Total labeled/total cells	Difference	Largest negative difference	Largest positive difference
Site 1 (n = 34)[b]	6.367[a] (2.718)	6.383 (2.713)	−0.015 (0.076)	−0.30	0.11
Site 2 (n = 35)[b]	6.330[a] (3.099)	6.337 (3.084)	−0.007 (0.082)	−0.20	0.22
Site 3 (n = 36)[b]	8.198 (4.035)	8.259[a] (4.004)	−0.062[c] (0.220)	−0.98	0.07
Site 4 (n = 34)[b]	7.930 (3.424)	7.934[a] (3.422)	−0.005 (0.049)	−0.13	0.13

[a] Method reported.
[b] No. of slides with scorable crypts.
[c] Larger mean difference is due to the inclusion of counts of single crypt columns and full crypts on the same slide.

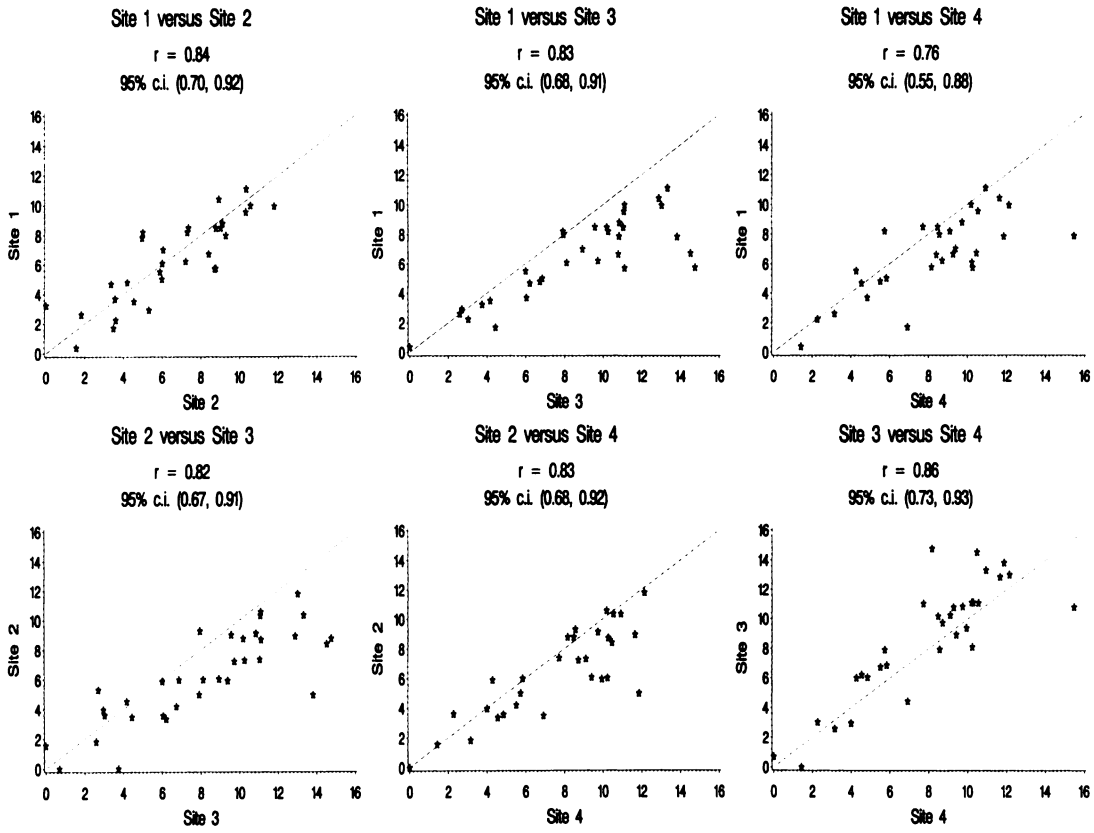

Fig. 1. Plots of the reported labeling indices for the pairs of sites. The observed labeling index with the line of identity is shown, with the estimated correlation coefficient and its 95% confidence interval. The correlations among the sites are high (all $r > 0.75$); however, note the differences in the magnitudes of the values reported by sites 3 and 4 relative to those reported by sites 1 and 2.

the magnitude of the crypt-to-crypt variability of the labeling index within a person is quite large compared with the biopsy-to-biopsy and even person-to-person variability (9, 10). Under these circumstances, computing the labeling index by dividing the number of labeled cells by the number of total cells across all crypts (the method of sites 3 and 4) may mask some of this crypt-to-crypt variability (3), particularly when crypt columns and full crypts are combined. Specifically, an estimate of the crypt-to-crypt variability within a subject is easily obtainable using the first method (SD of the observed crypt-labeling indexes), but such an estimate is generally not reported for the second method (although a weighted SD could be used). However, until we have a better understanding of the biological implications of within-subject labeling index variability, it may be difficult to distinguish between these two methods of labeling index computation.

It must be emphasized that these results may be somewhat idealized, because the technicians were told to count the same circled area on each slide. This restriction was imposed to narrow the focus to the reproducibility of the counting procedure, rather than the reproducibility of area selection and the counting procedure. One might reasonably expect that results obtained in the more common laboratory setting, in which the technician chooses the section of the slide to count, would likely show greater discrepancies and smaller correlations than those presented here. The procedures of all laboratories also restricted attention to crypts with normal morphology, which may obscure biologically relevant features (17, 18). Also, be-

cause our study was designed to assess the overall magnitude of interlaboratory reproducibility, we were unable to retrospectively identify the components of variation leading to the overall differences; such identification would require detailed mapping of the selected area, indicating which crypts were considered scorable and which cells were considered labeled. Finally, compartmental analyses to assess potential shifts in the proliferative zone were not performed. Without overall concordance in the number of scorable crypts and the reported overall labeling index, it was considered potentially misleading to perform secondary analyses, particularly because they would have poor statistical power (due to the absence of labeled cells within each compartment for most of the slides).

When this study was begun, BrdUrd had replaced tritiated thymidine in studies assessing cell proliferation. More recently, PCNA has been increasingly used (9, 12). A recent comparison of specimen preparation using both techniques suggested that BrdUrd labeling resulted in a larger proportion of unscorable specimens and a lower mean number of scorable crypts; PCNA labeling, therefore, was preferred (12). However, the considerations raised here regarding discrepancies among laboratories are even more applicable for PCNA, because it involves assessment of the level of staining intensity, rather than just the presence or absence of labeled cells, as in the case of BrdUrd. Therefore, it can be expected that differences among laboratories might be greater for PCNA compared with those seen here for BrdUrd. This possibility suggests that absorbance measurements of PCNA-staining intensity may be preferable to visual assessment.

Finally, these results underscore the advantages of procedure standardization across laboratories. Ongoing multicenter trials, therefore, are justified in the decision to use a single laboratory or to expend significant resources in laboratory standardization. Additionally, they emphasize that attempts to naively compare BrdUrd labeling index values reported by different laboratories could lead to spurious conclusions. However, criticism of the labeling index as an intermediate marker, based solely on problems of reproducibility, obscures the potential solvability of the technical issues involved. Here, we made no attempts to standardize procedures across laboratories or to suggest any changes from each laboratory's stated procedures. Therefore, in this setting, the high correlation among the results from the laboratories is reassuring. However, these results suggest that different laboratories may not only consider different cells labeled *versus* unlabeled, but they also may differ in the crypts they decide to count. These issues are the type that can be more easily addressed by the development of consensus working conferences (2), particularly if they include laboratory components. Our results underscore the necessity of such approaches, particularly as newer techniques, such as PCNA, become more prevalent.

Acknowledgments

We are indebted to Dr. Mikel Aickin for his helpful comments on an earlier version of the manuscript, two anonymous reviewers for their helpful suggestions, and Lois Bartz (University of Minnesota) for her technical support.

References

1. Kelloff, G. J., Malone, W. F., Boone, C. W, Steele, V. E., and Doody, L. A. Intermediate biomarkers of precancer and their application in chemoprevention. J. Cell. Biochem., *16G* (Suppl.): 15–21, 1992.

2. Lipkin, M. Summary of recommendations for colonic biomarker studies of candidate chemopreventive compounds in phase II clinical trials. J. Cell. Biochem., *19* (Suppl.): 94–98, 1994.

3. Aickin, M. If there is gold in the labeling index hills, are we digging in the right place? J. Cell. Biochem., *19* (Suppl.): 91–93, 1994.

4. Potten, C. S., Kellett, M., Roberts, S. A., Rew, D. A., and Wilson, G. D. Measurement of *in vivo* proliferation in human colorectal mucosa using bromodeoxyuridine. Gut, *33:* 71–78, 1992.

5. Alberts, D. S., Einspahr, J., Aickin, M., Hixson, L., Earnest, D., Roe, D., and Powell, M. Validation of proliferation indices as surrogate endpoint biomarkers. J. Cell. Biochem., *19* (Suppl.): 76–83, 1994.

6. Lipkin, M., Bhandari, M., Hakissian, M., Croll, W., and Wong, G. Surrogate endpoint biomarker assays in phase II chemoprevention clinical trials. J. Cell. Biochem., *19* (Suppl.): 47–54, 1994.

7. Bostick, R. M., Potter, J. D., Fosdick, L., Grambsch, P., Lampe, J. W., Wood, J. R., Louis, T. A., Ganz, R., and Grandits, G. Calcium and colorectal epithelial cell proliferation: a preliminary randomized, double-blinded, placebo-controlled clinical trial. J. Natl. Cancer Inst., *85:* 132–141, 1993.

8. Fireman, Z., Rozen, P., Fine, N., and Chetrit, A. Reproducibility studies and effects of bowel preparations on measurements of rectal epithelial proliferation. Cancer Lett., *45:* 59–64, 1989.

9. Einspahr, J., Alberts, D., Xie, T., Ritchie, J., Earnest, D., Hixson L., Powell, M., Roe, D., and Grogan, T. Comparison of proliferating cell nuclear antigen (PCNA) versus the more standard measures of rectal mucosal proliferation rates in subjects with a history of colorectal cancer and normal age-matched controls. Cancer Epidemiol., Biomarkers & Prev., *4:* 359–366, 1995.

10. Lyles, C. M., Sandler, R. S., Keku, T. O., Kupper, L. L., Millikan, R. C., Murray, S. C., Bangdiwala, S. I., and Ulshen, M. H. Reproducibility and variability of the rectal mucosal proliferation index using proliferating cell nuclear antigen immunohistochemistry. Cancer Epidemiol., Biomarkers & Prev., *3:* 597–605, 1994.

11. Bostick, R. M., Fosdick, L., Wood, J. R., Grambsch, P., Grandits, G. A., Lillemoe, T. J., Louis, T. A., and Potter, J. D. Calcium and colorectal epithelial cell proliferation in sporadic adenoma patients: a randomized, double-blinded, placebo-controlled clinical trial. J. Natl. Cancer Inst., *87:* 1307–1315, 1995.

12. Baron, J. A., Wargovich, M. J., Tosteson, T. D., Sandler, R., Haile, R., Summers, R., van Stolk, R., Rothstein, R., and Weiss, J. Epidemiological use of rectal proliferation measures. Cancer Epidemiol., Biomarkers & Prev., *4:* 57–61, 1995.

13. Fleiss, J. L. Statistical Methods for Rates and Proportions, Ed. 2, pp. 212–222. New York: John Wiley & Sons, 1981.

14. Rosner, B. Fundamentals of Biostatistics, Ed. 4, pp. 377–383,503–508. Boston: PWS-Kent, 1994.

15. Crowder, M. J., and Hand, D. J. Analysis of Repeated Measures, pp. 25–29. London: Chapman & Hall, 1990.

16. Miller, R. G., Jr. Simultaneous Statistical Inference, Ed. 2, pp. 67–70. New York: Springer-Verlag, 1981.

17. Araki, K., Ogata, T., Kobayashi, M., and Yatani, R. A morphological study on the histogenesis of human colorectal hyperplastic polyps. Gastroenterology, *109:* 1468–1474, 1995.

18. Savidge, T. C., Walker-Smith, J. A., and Phillips, A. D. Novel insights into human intestinal epithelial cell proliferation in health and disease using confocal microscopy. Gut, *36:* 369–374, 1995.

Reproducibility of the measurement of colonic proliferation using bromodeoxyuridine across laboratories.

D J Roe, D S Alberts, M J Wargovich, et al.

Cancer Epidemiol Biomarkers Prev 1996;5:349-353.