

An Evaluation of Rectal Mucosal Proliferation Measure Variability Sources in the Polyp Prevention Trial: Can We Detect Informative Differences among Individuals' Proliferation Measures Amid the Noise?¹

Lisa M. McShane,² Martin Kulldorff, Michael J. Wargovich, Cindy Woods, Madhu Purewal, Laurence S. Freedman, Donald K. Corle, Randall W. Burt, Donna J. Mateski, Michael Lawson, Elaine Lanza, Barbara O'Brien, William Lake, Jr., James Moler, and Arthur Schatzkin

National Cancer Institute, Bethesda, Maryland 20892 [L. M. M., A. S., L. S. F., D. K. C., E. L., M. K.]; M. D. Anderson Cancer Center, Houston, Texas 77030 [M. J. W., C. W., M. P.]; University of Utah, Salt Lake City, Utah 84132 [R. W. B.]; Kaiser Medical Center, Sacramento, California 95825 [M. L.]; Walter Reed Army Medical Center, Washington, DC 20307-5001 [D. J. M.]; Westat, Inc., Rockville, Maryland 20850-3129 [B. O.]; and Information Management Services, Rockville, Maryland 20852 [W. L., J. M.]

Abstract

We assessed components of total variability of bromodeoxyuridine (BrdUrd) and proliferating cell nuclear antigen (PCNA) assays of rectal mucosal proliferation in a subset of 390 participants from the U. S. National Cancer Institute's multicenter Polyp Prevention Trial. Biopsies were blindly double-scored by two technicians. For those participants for whom at least one evaluable biopsy was obtained, a mean of 2.0 and 2.6 biopsies, and 6.2 and 8.7 crypts/biopsy were evaluated, respectively, with the BrdUrd and PCNA assays. Factors such as clinical center, scorer, and month of biopsy collection significantly affected the observed values of the labeling index (LI) and proliferative height (PH). Therefore, it is essential to control or adjust for these variables in proliferation studies. Sources of random variation for LI and PH measures remaining after the aforementioned factors include between-participant variation and several sources of within-participant variation, including variation over time, between biopsies, and between multiple measurements on the same biopsy. Both LI and PH measurements exhibited substantial variability over time, between biopsies, and from reading-to-reading of the same biopsy. When other sources of

variability have been accounted for, the PCNA LI seems to have little between-participant variation. This brings into question its utility as a marker in colorectal cancer studies. The PCNA PH showed significant between-participant variability and may hold some promise as a useful marker in colorectal cancer studies. Results for BrdUrd were less conclusive. The BrdUrd LI showed marginally significant between-participant variation, whereas the corresponding variation for PH was nonsignificant.

Introduction

Measures of proliferative activity in rectal mucosal epithelial cells have received much attention as markers of risk for colorectal cancer and as potential surrogate end points in large bowel neoplasia prevention trials (1, 2). Immunohistochemical techniques can be used to label proliferating cells in rectal biopsy specimens. Two presently popular assays involve BrdUrd³ and PCNA labeling. Scoring of the processed biopsies involves recording, on a crypt-by-crypt basis, the positions of all labeled cells. A traditional measure of proliferation rate is the LI, defined as the number of labeled cells divided by the total number of cells. The LI may be computed on a biopsy level, crypt level, or crypt compartment level. Alternative measures that are designed to capture proliferative zone location include PH (3), defined as the mean position (in percentage) of labeled cells in the crypt where position is defined in terms of number of cells, ordered from base to lumen. Fig. 1 shows a diagram of a crypt with example calculations of the LI and PH.

Several studies examining the effect of various interventions on proliferative measures have yielded contradictory results, including two recent large randomized calcium supplementation trials (3, 4) and a recent randomized trial examining the effects of both wheat bran fiber and calcium supplementation (5). One of the calcium-only supplementation trials (3) showed no effect of calcium on rectal mucosal proliferation (PCNA and BrdUrd) either in terms of proliferative rate or proliferative zone location, whereas the second trial (4) also showed no effect of calcium on (PCNA) proliferative rate but did show a significant downward shift in proliferative zone in the calcium group. The third trial found that neither wheat bran fiber nor calcium supplements significantly reduced the [³H]thymidine LI in rectal mucosal crypts (total or compart-

Received 8/1/97; revised 4/17/98; accepted 4/21/98.

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 The majority of this research was conducted while L. M. M. was a Senior Staff Fellow in the Biometry Branch in the Division of Cancer Prevention and Control at the National Cancer Institute. The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the Department of the Army or the Department of Defense.

² To whom requests for reprints should be addressed, at National Cancer Institute, Biometric Research Branch, CTEP, DCTD, Executive Plaza North, Room 739, 6130 Executive Boulevard MSC 7434, Bethesda, MD 20892-7434.

³ The abbreviations used are: BrdUrd, bromodeoxyuridine; IES, Intermediate Endpoint Study; LI, labeling index; PCNA, proliferating cell nuclear antigen; PH, proliferative height; PPT, Polyp Prevention Trial; MDACC, M. D. Anderson Cancer Center; ICC, intraclass correlation coefficient.

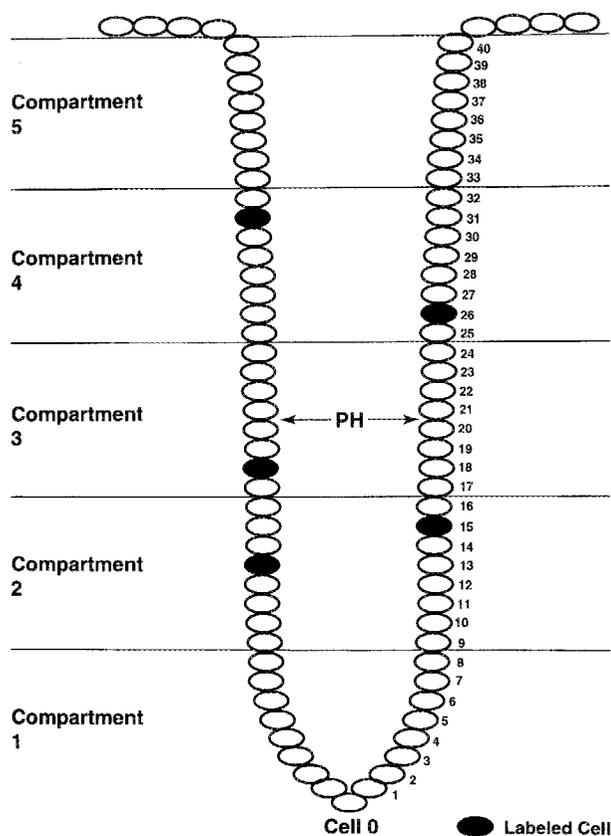


Fig. 1. Diagram of a typical rectal mucosal crypt cross-section. In the example shown, the LI is 6.2% [(5 labeled cells/80 total cells) \times 100%], and the PH is 51.5% [mean position of labeled cells \times 1/height = (31 + 26 + 18 + 15 + 13)/5 \times 1/40 \times 100%].

mental analysis). There has been speculation that the apparently conflicting results of such studies may be due at least in part to high variability ("noise") in the proliferative measures. High variability would tend to attenuate interindividual differences and would make it difficult to demonstrate real intervention effects. Knowledge of the variability in the proliferative measures is, therefore, essential in designing and interpreting studies of rectal mucosal proliferation.

In this study, we assess the reproducibility of two proliferation measures and the biopsy-level LI and PH obtained from both the BrdUrd and PCNA assays, and we investigate the degree of between-person variability in these measures, using extensive data from the U. S. National Cancer Institute's PPT (6). For each of these four measure-assay combinations, we identify and estimate source-specific components of the total variability and discuss implications for future trials using these proliferative measures as outcomes.

Materials and Methods

Subjects. Rectal mucosal specimens were obtained from 392 subjects participating in the IES within the PPT (6). The PPT is an ongoing multicenter randomized dietary intervention study of the effect of a low fat, high fiber, high fruit and vegetable eating plan on large bowel adenomatous polyp recurrence. The 2079 participants, 35 years of age or older with a recent history of adenoma removal, were randomized to either the dietary

counseling group (low fat, high fiber, fruits, and vegetables) or the control group (no dietary counseling) (7). The IES was conducted at three of the eight PPT clinical centers [Kaiser-Oakland (Sacramento), Walter Reed, and Utah] and it required separate informed consent. For the IES participants, eight mucosal biopsy specimens are collected from the large bowel on three separate occasions: shortly after the baseline examination (T_0); at the end of the first year (T_1); and at the end of the fourth year (T_4). To date, only the T_0 and T_1 examinations have been completed for all study subjects, and the analyses presented in this study are based on the data collected only at those two visits. Due to some delays in entering subjects into the IES, some patients were able to participate at T_1 but not at T_0 , and there was some dropout between T_0 and T_1 examinations. On each occasion, three of the biopsy specimens are assayed for cell proliferation by the BrdUrd technique, three are assayed by the PCNA technique, and two are frozen for future analysis.

Biopsy Preparation. Biopsies of the rectal mucosa were obtained from consenting individuals who were eligible participants in the IES substudy of the PPT. The protocol for bowel preparation was not universally standardized among centers as they were allowed the flexibility to prepare their patients in keeping with local customs. For the vast majority of the biopsy collections taken during colonoscopy the bowel preparation was Co-lyte solution or Golytely. We have shown in preliminary studies at the MDACC that the Co-lyte prep did not significantly alter proliferation kinetics when compared with biopsies taken without a bowel prep; thus, we are confident that Co-lyte did not affect BrdUrd and PCNA values. Samples taken during the flexible sigmoidoscopy procedure or a procedure for biopsy acquisition exclusive of colonoscopy were generally taken from an unprepared colon. The six biopsies (three each for BrdUrd and PCNA) were taken from each participant at the same time. The biopsies were carefully removed from the endoscopy forceps and quickly placed on a strip of bibulous paper, immersed in MEM (Sigma Chemical Co., St. Louis, MO) and transported within 15 min to an area where the orientation procedure was possible. The biopsies were then oriented with the aid of a dissection microscope on the paper strips to assure maximum exposure to the medium containing BrdUrd or fixative.

BrdUrd Assay. The BrdUrd assay was conducted in disposable borosilicate sample vials (Fisher Scientific Co., Pittsburgh, PA) under hyperbaric oxygen. Well-oriented biopsies on bibulous paper were placed in the sample vial with MEM containing 50 μ M BrdUrd (Sigma Chemical Co.). Approximately 2 ml of 95% O_2 /5% CO_2 was injected into the tube before incubation through the screw-top septum. The biopsies were incubated for 1 h at 37°C, with agitation. After the hour incubation the medium was carefully decanted from the sample vial and the tube was gently filled with 70% ethanol. The biopsies were batched and shipped to a central repository, then shipped to the MDACC for analysis. The shipping and storing of all specimens was carried out under the direction of the trial's Data and Nutrition Coordinating Center (Westat, Inc., Rockville, MD).

PCNA Assay. For the three biopsies for the PCNA assay a similar orientation step was followed, but the transport medium was decanted and the shipping tube was filled with 70% ethanol. Biopsies were similarly shipped to the central repository, then reshipped to the MDACC for analysis.

Intermediate Endpoint Analysis. Analysis of the two biomarker assays consisted of an early quality control assessment followed by immunohistochemistry for BrdUrd or PCNA.

scoring of labeled cells, and data entry. Batched samples were received from the repository and immediately checked for damage in transit or missing samples, and logged into a tracking database. All biopsies were processed for histological section and imbedded in paraffin. Sections of thickness 4- μ m were cut from the samples and placed on poly-l-lysine coated slides. A quality control assessment was performed by examining the unstained slides under light microscopy for the presence of sufficient numbers of well-oriented crypts before immunohistochemistry was performed.

Sections of rectal biopsies with adequately well-oriented crypts were immunostained for BrdUrd using an anti-BrdUrd monoclonal antibody (Becton Dickinson, San Jose, CA) or anti-PCNA (PC-10 clone; Signet Laboratories, Inc., Dedham, MA) for the PCNA assay. Exposure to the monoclonal antibodies was assisted by the use of a semi-automated Sequenza device (Scimetric, Inc., Missouri City, TX). Visualization of BrdUrd or PCNA-labeled cells was achieved by using the immunoperoxidase method with diaminobenzidine as the chromagen.

Several steps were taken throughout the process to assure blinding. All batched samples were received from the repository with a coded descriptor to allow knowledge only of which participating center contributed the samples. This facilitated a discussion with the participating centers regarding biopsy quality if problems occurred. Slides were randomly assigned among the five members of the lab staff so that each slide was scored independently by two observers. If a wide discrepancy occurred in their assessments the biopsies were rescored using a third scorer as the arbiter. Well-stained and -oriented specimens of human colon mucosa were always included in batched slides to be scored as positive controls.

Only well-oriented crypts were scored. These were defined as a crypt for which the base touched the muscularis mucosa and for which a U-shaped pattern could be traced with an open lumen at the apex of the crypt. Each scorer was well-trained to recognize acceptable, marginal, and unacceptable crypts as well as acceptable staining patterns. Crypt selection and scoring were performed according to an agreed protocol. The scorer oriented the slide on the microscope stage, and from left to right scored successive crypts and successive levels of the biopsy until all scorable crypts were enumerated. If all three biopsies for each assay failed to yield at least eight scorable crypts, new sections were recut from the block in an attempt to make up the deficit. For each crypt, a crypt height was determined by identifying the center cell at the crypt base and counting the ordinal number of unstained and stained cells along the crypt axis. For BrdUrd, only the deepest stained cells in the crypt were called "labeled" whereas for PCNA, the darkest labeled cells and the next lighter shade of brown stained in the crypt were called "positively labeled." To promote standardization of scorings, each scorer was supplied with a manual containing photographs of optimally stained human colon crypts with PCNA from which the group had decided what constituted a dark labeled PCNA cell, and its next lightest grade. Biopsies containing crypts with acceptable orientation and staining pattern as described above were called "evaluable."

All data resulting from the proliferation analysis were recorded by hand into hardcopy binders. The data sheets were copied and batch shipped to the coordinating center for double data entry and subsequent analysis.

Statistical Methods. Differences in proliferation measures can be attributed to many sources. We envision that each person has

a 'true' underlying mean proliferation measure. An effective treatment causes a systematic shift in mean proliferation levels for many individuals. In addition to any systematic shifts due to treatment effects, there are differences in mean levels across participants within a treatment group. Some of these differences may be attributable to identifiable factors such as clinical center, whereas other differences cannot be explained. We refer to this portion of the variation as the between-participant variation.

There are several sources of within-participant variation. Multiple measurements made on the same individual will fluctuate about that individual's true mean. There may be differences in the measurements obtained from multiple biopsies collected at the same time from the same individual, and there may be differences in measurements over time in the same individual, over and above the magnitude of differences expected due to different biopsies. Some of the differences over time might be explained by identifiable factors such as, for example, time of day or season of year of biopsy collection; or, participating in a trial may modify behavior, such as physical exercise, that could affect proliferative activity over time, regardless of the treatment group assignment. Some portion of the variation over time is likely to remain unexplained, and we will refer to this random portion as the within-participant variation over time. This variation could be due to biological fluctuations in the participant, or to some random changes in biopsy collection or processing affecting all biopsies collected from a participant at a particular visit. Multiple biopsies collected at the same time may yield different proliferation measurements due to factors including variability in underlying proliferation processes in adjacent tissues or due to random differences in biopsy handling. We refer to this type of variation as biopsy-to-biopsy variation. Lastly, multiple readings obtained on a single biopsy specimen may differ due to measurement error. In scoring a biopsy, each scorer must exercise his/her judgement to select scorable crypts and to determine the number and location of labeled and unlabeled cells. Besides the systematic differences between scorers, these subjective judgements may differ randomly between scorers presented with the same biopsy slide (interscorer variability), or they may differ randomly between two occasions when the same scorer judges a particular slide (intrascorer variability). It is this inter- and intrascorer variability that we refer to as measurement error.

All biopsies in our study were blindly scored by two scorers chosen at random from five members of the lab staff. Although crypt-level scoring was performed by each scorer, two different scorers for any one biopsy may not have judged the same subset of crypts to be scorable. Because crypts were not labeled with identifiers, it was not feasible to match individual crypt scorings. Hence, all of the proliferative measures are summarized to the biopsy level.

To analyze the data, we developed a mixed model, defined as a model containing both fixed and random effects (8). Fixed effects are those attributable to a factor with a finite number of levels that either we control experimentally or we adjust for, for example, treatment and scorer effects in our study. In contrast, random effects are those attributable to a factor whose levels are regarded as a random sample from some larger population, for example, participants or biopsies in our study. If we were to replicate the study, we would have a different collection of participants and biopsies, but the same treatments and scorers. Our mixed model contains fixed effects for treatment, follow-up visit (T_0 and T_1), clinical center, hour and month of biopsy collection, scorer, scorer-specific effects of calendar time of scoring, and two-way interactions between scorers and

visits and between visits and treatments; and random effects for between-participant variation, within-participant variation over time, biopsy-to-biopsy variation, and measurement error. Initially, fixed effects for age, race, gender and additional interaction terms were incorporated into the model. Occasionally one of these was found to be significant for one of the four proliferation outcome measures, but their inclusion in the model did not substantially affect variance component estimates or other fixed effect estimates, so for simplicity and consistency we did not include them in our final models.

The usual assumptions for mixed models are that the random effects are independent, normally distributed random variables with zero means, and their variances are referred to as "variance components." These variance components describe the contributions of each of the sources of random variability to the total variability in the measurements after adjusting for all fixed effects, and are of prime interest in this study. We let $\sigma_{participant}^2$, σ_{time}^2 , σ_{biopsy}^2 , and σ_{error}^2 denote the variance components associated with the between-participant variation, within-participant variation over time, biopsy-to-biopsy variation, and measurement error, respectively. The variance inherent in a single measurement (adjusted for all fixed effects) made on a single biopsy obtained from a single participant on a single occasion is equal to the sum of all of the variance components, namely $\sigma_{participant}^2 + \sigma_{time}^2 + \sigma_{biopsy}^2 + \sigma_{error}^2$. This model assumes constant variance; that is, the magnitude of the total variance and each of its components does not depend on additional factors such as an individual's true underlying proliferative measure or other participant-specific characteristics. For example, the between-participant variation is the same whether one considers the participants examined at T_0 or at T_1 , or in the intervention or control group; the biopsy-to-biopsy variation is the same for all participants.

To check the model assumption of normality, we examined the data on both untransformed and log transformed scales using quantile plots and histograms. Log LI and log PH seemed to be reasonably approximated as normally distributed. Lipkin's Φ_h (9) also was computed but did not seem normally distributed for our data. Hence, we could not apply these methods of variance components estimation. Assessing the variability in Φ_h would require development of new statistical methods that would be beyond the scope of this study.

Plots of the log-transformed proliferative measures versus total number of cells counted suggested mild departures from the assumption of constant variance, although most authors have treated these measures as having constant variance. To investigate the sensitivity of our analysis to the assumption of constant variance, we also performed a weighted mixed model analysis using techniques proposed by Grambsch *et al.* (10) that allow the variance of the proliferative measure to depend on its mean and on the total numbers of cells counted in the biopsy. We fitted mixed models (both weighted and unweighted) to the log transformed data for each of the four measure-assay combinations using the restricted maximum likelihood method in SAS PROC MIXED (11). This program yields estimates of the individual variance components along with their estimated SDs, from which we computed approximate confidence intervals and performed significance tests. We found that the variance components estimated under the weighted mixed models were similar to those under the constant variance models. Thus, for simplicity, we report the estimates obtained assuming constant variance.

The variance component estimates can be used to estimate the total variance in the summary measure obtained for an individual at a particular point in time, and to estimate the ICC

Table 1 Numbers of participants from whom evaluable biopsy results were obtained

	Baseline (T_0)	1 Year (T_1)	Both T_0 and T_1
BrdUrd			
Control	99	127	71
Intervention	115	142	77
Total	214	269	148
PCNA			
Control	155	166	133
Intervention	159	172	134
Total	314	338	267

which measures the ability of a measurement technique to distinguish between individuals' true marker values. In our study, the proliferative measure for an individual at a particular time usually is obtained as the mean of two scorings of each of three biopsies. In general, the variance of the mean of readings taken on several biopsies (B), with each biopsy scored by several independent scorers (S), obtained from an individual on a single occasion, is

$$\sigma_{total}^2 = \sigma_{participant}^2 + \sigma_{time}^2 + \sigma_{biopsy}^2/B + \sigma_{error}^2/BS. \quad (A)$$

The divisors of B and BS on the biopsy and error variances, respectively, reflect the reduction in variance due to averaging over the multiple biopsies and scorings. The magnitude of the square root of the total variance (A) relative to the size of treatment or intervention effect one wishes to detect is a determinant of sample size for trials. The ICC describes the percentage of the total variance due to between-participant variation, and is defined as

$$ICC = (\sigma_{participant}^2 / \sigma_{total}^2) \times 100\%. \quad (B)$$

The ICC lies between 0 and 100%. A value close to 100% indicates a highly reproducible assay, and a value close to 0% indicates an assay that is unable to distinguish among individuals. The absolute magnitude of $\sigma_{participant}^2$ should be considered in addition to the ICC, for it is possible to have an assay that is highly reproducible, yet the true levels of the marker that the assay is measuring barely differ between individuals. In this situation, the ICC could approach 100%, but the marker may have little scientific interest unless extremely small differences in marker levels translate to important differences in other variables or outcomes of interest.

Interscorer reliability also can be estimated from the variance components model. It is given by $r = (\sigma_{participant}^2 + \sigma_{time}^2 + \sigma_{biopsy}^2) / \sigma_{total}^2$, and represents the correlation between two randomly chosen scorers' measurements of the same biopsy. Our data did not allow us to estimate intrascorer reliability because no scorer ever scored the same biopsy more than once.

Results

Of the 392 patients from whom biopsy specimens were collected in this study, all but 2 had evaluable results on at least one biopsy. The numbers of participants for whom we have obtained evaluable results, categorized by BrdUrd and PCNA assay, and by visit, are presented in Table 1.

Only some of the participants had results at both baseline (T_0) and 1-year (T_1) examinations. For BrdUrd, 148 participants had results at both T_0 and T_1 , compared with 66 participants with results at T_0 only and 121 participants with results at T_1 only. For PCNA, 267 participants had results at both T_0

Table 2 LI least squares means^a for levels of selected fixed effects and associated *P*-values for tests of significance^b

	BrdUrd	PCNA
Clinical center		
1	4.4	5.7
2	4.2	3.9
3	3.8	4.7
	<i>P</i> = 0.027	<i>P</i> < 0.0001
Scorer		
1	3.8	3.8
2	3.7	3.9
3	6.0	6.7
4	4.2	5.7
5	3.5	4.0
	<i>P</i> = 0.015	<i>P</i> = 0.0020
Hour of biopsy		
7:00 a.m.–8:59 a.m.	4.1	4.9
9:00 a.m.–9:59 a.m.	3.9	4.7
10:00 a.m.–10:59 a.m.	4.2	4.5
11:00 a.m.–12:59 p.m.	4.2	4.8
1:00 p.m.–7:00 p.m.	4.2	4.7
	<i>P</i> = 0.34	<i>P</i> = 0.17
Month of biopsy		
January	4.4	5.2
February	4.0	5.4
March	3.8	4.3
April	3.6	3.9
May	4.3	4.2
June	4.1	4.6
July	4.5	4.5
August	4.2	4.4
September	3.8	5.2
October	4.5	4.6
November	4.2	5.2
December	4.2	5.2
	<i>P</i> = 0.22	<i>P</i> < 0.0001

^a Back-transformed to original scale from least squares means computed on log scale.

^b Based on approximate *F*-tests as computed by SAS PROC MIXED.

and T₁, compared with 47 participants with results at T₀ only and 71 participants with results at T₁ only. Due to technical difficulties with the BrdUrd assay, PCNA results were successfully obtained among a greater proportion of participants than BrdUrd results, and this explains the larger number of participants under PCNA. Evaluable results were obtained on the BrdUrd assay for 42% of the available T₀ biopsies and for 55% of the T₁ biopsies. For PCNA, the evaluable percentages were 83% at T₀ and 84% at T₁. These evaluable biopsy rates translate to collection of at least one evaluable biopsy on BrdUrd for 66 and 78% of participants at T₀ and T₁, respectively, and for 95 and 97% of participants on PCNA at T₀ and T₁, respectively. Considering all participant visits in which at least one evaluable biopsy was obtained, we obtained an average of 2.0 evaluable biopsies/participant-visit using the BrdUrd assay and 2.6 biopsies/participant-visit using the PCNA assay. The average number of crypts scored/biopsy was 6.2 for BrdUrd and 8.7 for PCNA.

Tables 2 and 3 show, for the LI and PH, respectively, least squares means (transformed back to original scale) and tests of significance associated with selected fixed effects (effects involving treatment are not presented to ensure confidentiality of results in the ongoing trial). The distribution of patients to clinical centers was approximately 25, 30, and 45% for centers 1, 2, and 3, respectively. Scorers 3 and 4, combined, scored fewer than 12% of all biopsies. Scorers 1, 2, and 5 each scored approximately one-third of the

Table 3 PH least squares means^a for levels of selected fixed effects and associated *P*-values for tests of significance^b

	BrdUrd	PCNA
Clinical center		
1	25	27
2	28	26
3	27	26
	<i>P</i> = 0.0009	<i>P</i> = 0.55
Scorer		
1	27	26
2	28	27
3	29	27
4	22	25
5	28	27
	<i>P</i> < 0.0001	<i>P</i> < 0.0001
Hour of biopsy		
7:00 a.m.–8:59 a.m.	26	26
9:00 a.m.–9:59 a.m.	26	26
10:00 a.m.–10:59 a.m.	26	26
11:00 a.m.–12:59 p.m.	27	26
1:00 p.m.–7:00 p.m.	29	27
	<i>P</i> = 0.15	<i>P</i> = 0.90
Month of biopsy		
January	28	27
February	25	28
March	25	26
April	26	26
May	26	25
June	28	26
July	26	26
August	28	26
September	27	27
October	28	27
November	26	26
December	27	26
	<i>P</i> = 0.041	<i>P</i> = 0.0045

^a Back-transformed to original scale from least squares means computed on log scale.

^b Based on approximate *F*-tests as computed by SAS PROC MIXED.

remaining biopsies. Approximately 75% of all biopsies were collected before 11 a.m., at times evenly distributed throughout the morning. Approximately 10 and 15% of biopsies were collected from 11 a.m.–1 p.m., and after 1 p.m., respectively. The percentages of biopsies collected each month varied from 5–12%, with greater numbers collected in May/June and fewer collected in the summer. Scorer effects were strongly significant for both proliferation measures using both assays. For both BrdUrd and PCNA, scorers 1, 2, and 5 tended to report lower LIs than scorers 3 and 4; scorers 1 and 4 tended to report lower PHs. Clinical center effects were significant for all measures except PH by PCNA. The ordering of the clinical center effects for the LI was not consistent between BrdUrd and PCNA. Before adjusting for clinical center differences, the hour of biopsy collection seemed to have a significant effect; however, it was discovered that the distribution of biopsy collection times differed substantially between clinical centers. Clinic 1 collected biopsies throughout the day, clinic 2 collected biopsies mid-morning, and clinic 3 collected a large proportion of the biopsies in early morning. When both clinic effects and time of biopsy collection were included in the model, the hour of biopsy collection was nonsignificant whereas clinic effects remained significant. The month of biopsy collection was significant for all measures except LI for BrdUrd. The monthly pattern was not strongly consistent between BrdUrd

Table 4 Variance components analysis

	Participant	Time	Biopsy	Error	ICC ^a (%)
Log LI					
BrdUrd					
Var (SE) ^b	0.0215 (0.0126)	0.0464 (0.0143)	0.1000 (0.0082)	0.0577 (0.0027)	19.4 ^c
<i>P</i> ^c	0.089	0.0012	<0.0001	<0.0001	
Range ^d	(3.0-5.4) ^e	(2.6-6.2)	(2.1-7.5)	(2.5-6.5)	
PCNA					
Var (SE)	0.0059 (0.0059)	0.0598 (0.0077)	0.0262 (0.0034)	0.0867 (0.0030)	6.6 ^c
<i>P</i>	0.32	<0.0001	<0.0001	<0.0001	
Range	(3.4-4.7) ^e	(2.5-6.5)	(2.9-5.5)	(2.2-7.2)	
Log PH					
BrdUrd					
Var (SE)	0.0021 (0.0032)	0.0114 (0.0045)	0.0353 (0.0033)	0.0262 (0.0013)	7.1 ^c
<i>P</i>	0.52	0.010	<0.0001	<0.0001	
Range	(26-31) ^e	(23-35)	(19-41)	(20-39)	
PCNA					
Var (SE)	0.0048 (0.0013)	0.0063 (0.0014)	0.0129 (0.0010)	0.0193 (0.0007)	25.8
<i>P</i>	0.0003	<0.0001	<0.0001	<0.0001	
Range	(24-32)	(24-33)	(22-35)	(21-37)	

^a ICC for a mean proliferative measure comprised of three biopsies scored by two independent scorers, computed using expressions (A) and (B).

^b Restricted maximum likelihood variance component estimate (var) and its SE based on natural log transformed proliferative measures.

^c Reported *P* values are those supplied by SAS PROC MIXED, and they are based on assuming that the Wald test statistics have approximate standard normal distributions, although this approximation may not be very accurate since we are testing on the boundary of the parameter space (variance components are constrained to be greater than zero). The likelihood ratio tests have the same limitation (23), and yield *P* values very similar to the Wald *P* values reported above. For the case of testing a single variance component, an approximate correction that has been suggested is to divide the likelihood ratio *P* value by 2 (24, 25). With this correction, the participant-to-participant variance component for BrdUrd log LI becomes marginally significant (*P* = 0.045). The participant-to-participant variance component remains highly significant for PCNA log LI, and nonsignificant for both PCNA log LI and BrdUrd log PH. All other variance components remain highly significant for all four proliferation measures.

^d Estimated range of variation in proliferative measure, on original scale, due to that variance component alone. Calculated as $mean \times \exp(\pm 2\sqrt{\text{variance component}})$ where $mean = 4\%$ for LI, and $mean = 28\%$ for PH.

^e Between-participant variance is not significantly different from zero.

and PCNA. Also, not presented in Tables 2 and 3, there was evidence of significant drift over time by some of the scorers on some of the proliferation measures.

Table 4 shows the estimated components of variance and their SEs, on the log scale, for the two proliferation measures using both BrdUrd and PCNA assays. These variance components estimate the random variation remaining after adjusting for the fixed effects. These variance component estimates were obtained from the combined treatment and control groups to obtain a more precise estimate than using the control group alone. Because treatment group by time interactions were removed as part of the fixed effects portion of the model, treatment effects, if present, should not be inflating within person variability estimates. Below the variance component estimates are the *P* values associated with a test that the variance component is zero. All variance components were significantly different from zero except the between-participant variances for the LI using PCNA and both the LI and PH assayed with BrdUrd. The between-participant variance for the LI assayed with BrdUrd approached significance. In brackets are the 2.5 and 97.5 percentiles, transformed back to the original scale, that would result from these individual variance estimates, assuming a mean LI of 4% and mean PH of 28%. These intervals indicate the estimated range of variation in the measure due to the individual variance sources. The reported ICC is an estimate of the reproducibility of the assay for determining a subject's true underlying mean proliferative measure in our trial in which, at each participant's visit, three biopsies are scored by two independent scorers. Interscorer reliabilities (*r*) of the LI were estimated as 0.74 and 0.51 for BrdUrd and PCNA, respectively. For PH, the interscorer reliabilities were 0.65 (BrdUrd) and 0.55 (PCNA).

Discussion

Our study is one of the largest reported cell proliferation studies that examines variability in proliferative measures. It involved evaluable biopsies from 390 participants compared with other studies examining variability in BrdUrd, PCNA, tritiated thymidine, or whole crypt mitotic count proliferation assays based on biopsies from 21 or fewer subjects (10, 12-17). We performed a variance components analysis on the LI and PH, but for reasons of non-normally distributed data, we could not apply these statistical methods to Φ_h . Nonetheless, we felt that PH was an appealing alternative to Φ_h in that it uses even more information about labeled cell height than Φ_h because it is an average of relative crypt heights of labeled cells, whereas Φ_h is an average of binary indicators of upper crypt occupancy. Although we cannot estimate interpretable variance components for Φ_h using the methods in this study that rely on an assumption of normally distributed data, Φ_h will be examined for treatment effects at the conclusion of the trial. We begin with a discussion of some reasons why our results may differ somewhat from those obtained in previous studies, and we follow with a discussion of the findings that will have broad implications for design and analysis of future prevention trials using rectal mucosal cell proliferation measures as surrogate end points.

Even if the variance components models are identically specified in two different variability studies, there are many reasons why the individual variance component estimates could differ between studies. First, variance component estimates themselves tend to have large variability and require large sample sizes to be estimated with substantial precision. Second, even for large studies in which variance components can be estimated precisely, the estimates may differ due to differences

in study populations, bowel preparation, biopsy collection methods, and scoring techniques. The importance of standardization has been recognized (18, 19), and the results presented here should be interpreted with this in mind. Standardization will significantly enhance the ability to effectively plan studies and compare results across studies.

Our separation of the total variance into components differs somewhat from the separation used by the authors mentioned previously (10, 12–17). In estimating the variance components, many of these authors separately estimated crypt-to-crypt variance components and biopsy-to-biopsy variance components, whereas for reasons having to do with the blinded double scoring discussed previously, we could not estimate a separate crypt variance component. The crypt-to-crypt variability is not missing from our total variance; it is absorbed partly into the biopsy-to-biopsy variance and partly into the measurement error variance. Our measurement error also incorporates some intrascorer variability which is absorbed into the crypt-to-crypt variability of the models used by the other authors. If the number of scored crypts/biopsy and number of scorers is comparable, then models with and without separate crypt variance components should give similar estimates of the total variability. One feature of our model that has not been present in variance components models presented by other authors is a term representing within-participant variation over time. Other variance components analyses typically have been based on data collected at a single point in time, and estimated between-participant variances are inflated by within-participant variation over time. Because our data were collected at two time points, 1 year apart, we could estimate both a between-participant variance (representing variation between participants' true long-term mean levels) and a within-participant over time variance.

Our variance component analyses allow us to measure the strength of the participant-specific "signal" amid within-participant noise comprised of random temporal fluctuations, biopsy variability and measurement error. After adjustment for fixed effect factors, our results indicate that there remains significant between-participant variation only for the PH using the PCNA assay. The between-participant variation approaches significance for the LI using the BrdUrd assay. However, even for the proliferation measures having at least marginally significant between-participant variation, it is noticeable that this variation is much smaller than the variation within a participant. This is reflected in the fact that the ICCs shown in the final column of Table 4 are generally low (not significantly different from zero at the 0.05 level for three of the four measures), indicating that these BrdUrd and PCNA proliferation measures are noisy measurements from which it is difficult to extract a strong "participant-specific signal." Another noteworthy feature of Table 4 is that the estimated biopsy-to-biopsy variances for the LI and PH using BrdUrd are higher than the corresponding estimates for the PCNA assay. This may be a consequence of the greater difficulty of preparing BrdUrd biopsy specimens as compared with PCNA (20). Also, the measurement error variance associated with PCNA LI is higher than that for BrdUrd, possibly due to a greater degree of subjective judgment required to distinguish between multiple staining intensity levels. However, the measurement error variance associated with PCNA PH is lower than that for BrdUrd. This may reflect the larger average number of labeled cells using PCNA.

Our 0.74 interscorer reliability estimate for the BrdUrd LI was similar to the 0.79 figure reported by Bostick *et al.* (21). Our interscorer reliability estimate for PCNA LI was 0.51 compared with 0.92 reported by Bostick *et al.* using both only darkest labeled cells or all labeled cells. This disparity for the PCNA LI could indicate that our scoring method, which in-

cluded darkest and next darkest stained cells, may introduce more variability than using only darkest or all stained cells. It could also simply result from the fact that because our study used more scorers than most studies, we had a greater chance of observing a higher degree of between-scorer variability.

The significant effects of clinical center, scorer, month of biopsy collection, and drifts in the measurements over time for most of the proliferation measures emphasize the need to collect information on these factors, control for them in the design, and/or adjust for them in the analyses. If not accounted for, the effects of these factors could inflate the between participant variance component (as well as the ICC) so that what would seem to be "true" differences between individuals may really be differences in, say, biopsy month or clinical center. Other components of the variability may be affected as well and result in the need for increased study sample sizes. Most importantly, if these factors are not balanced across treatment groups and not adjusted for in the analysis, treatment group comparisons will be biased.

Although our study did not find a significant effect of hour of biopsy collection, there have been some reports of diurnal variation in rectal mucosal proliferation (22). Due to the strong association between time of biopsy collection and clinical center in our study and the fact that we did not control for time of amount of last food intake, our study may have had difficulty in detecting diurnal variation. A study specifically designed to address diurnal variation would be of interest.

The strong effect of clinical center potentially has at least three component sources: (a) the effect may be due in part to procedural differences among the clinical centers, including bowel preparation and biopsy collection, handling, and processing variations; (b) clinic differences could be reflecting environmental or life-style differences between the populations from which the clinics are drawing patients, yet to have no between-participant variation remaining within clinics would be quite surprising; and (c) the possibility that there exists some unknown biological characteristic, either cell proliferation itself or some other biological characteristic highly correlated with it, that differs between clinical center populations but is highly homogeneous within each population. Given the diversity of individuals within each clinical center population, it would be difficult to conceive of such a characteristic. Hence, although one could argue that by adjusting for clinical center we could be somewhat underestimating between-participant variation, the important point is that even within a clinical center population we would expect that if a marker were informative, it would show some variation.

From a cancer biology perspective, our findings of non-significant between-participant variation in three of the four proliferation measures cast doubt on the utility of some of these measures as predictive markers for the development of adenomatous polyps and possibly colorectal cancer in this patient population. The between-participant variation for the PCNA LI, adjusted for other fixed effect factors, was extremely small and not statistically significantly different from zero. Potentially, if we had an even larger sample size, the between-participant variation would have reached statistical significance. Nonetheless, the modest SE suggests that the between-participant variation, if any, is quite small. The interpretation of this result is that individuals in this patient population have generally the similar PCNA LIs, yet we anticipate 30–40% will develop adenomas over the 4 years of the study, and the remaining 60–70% will not. This raises serious questions about this marker's ability to predict future adenoma development in patients who had previously developed adenomatous polyps. However, these results cannot address whether or not this marker may be

useful for predicting polyp development among individuals without a history of polyps. Because the incidence of colorectal cancer in this population will be very small, it would be difficult to detect a relationship between colorectal cancer and cell proliferation directly. But, because it is believed that a large portion of colorectal cancers arise from adenomatous polyps, it would be surprising to find a good predictive marker for colon cancer that showed virtually no predictive ability for adenomas. The between-participant variation in the BrdUrd LI only approached statistical significance, but it was more than three times as large as the corresponding variation for PCNA. The lack of statistical significance could be due to reduced samples sizes (power) for BrdUrd, but the decidedly minimal between-participant variation for PCNA does not provide encouragement for the analogous BrdUrd measure. The between-participant variation for PCNA PH was highly significant, and the accompanying ICC was 25.8%. Therefore, PCNA PH may hold some promise as a useful colorectal cancer marker. However, between-participant variation for BrdUrd PH was not significant, and the estimate was less than half that for the PCNA. It also should be noted that if bowel preparation had been standardized across participants, the between-participant variation estimates might have been even smaller.

One might reasonably expect that treatment effects would have magnitudes no greater than the range of normal variation in polyp patients. Therefore, if any of these four measures turn out to be useful, the small magnitude of the between-participant variation for all of them suggests that effective treatments may result in very small changes in these proliferation measures. The feasibility of conducting trials to detect very small treatment effects depends on the size of those effects relative to the total variance given by expression (1). For PCNA PH, trials would be manageable. Approximately 150 subjects in each of the treatment and control groups would be required to have 90% power to detect a 5% decrease in PH assuming that for each participant, three biopsies are scored by each of two scorers, and all other fixed effect factors are controlled. Whether such small changes will be of clinical interest ultimately depends on the relative importance of cell proliferation in adenomatous polyp development and colorectal carcinogenesis, specifically, what magnitude of decrease in cell proliferation, if any, corresponds to a meaningful decrease in adenoma and colorectal cancer incidence. A partial answer to this question may eventually emerge from PPT, by relating cell proliferation rates to the recurrence of adenomas among the participants.

Acknowledgments

We thank J. Walter Kikendall for clinical assistance; and Phil Gray, Richard Terrell, and Kathryn Wergen for technical assistance in conducting the laboratory assays.

References

- Einspahr, J. G., Alberts, D. S., Gapstur, S. M., Bostick, R. M., Emerson, S. S., and Gerner, E. W. Surrogate end-point biomarkers as measures of colon cancer risk and their use in cancer chemoprevention trials. *Cancer Epidemiol. Biomark. Prev.*, 6: 37-48, 1997.
- Schatzkin, A., Freedman, L. S., Dorgan, J., McShane, L., Schiffman, M., and Dawsey, S. M. Surrogate endpoints in cancer research: a critique (Editorial). *Cancer Epidemiol. Biomark. Prev.*, 5: 947-953, 1996.
- Baron, J. A., Tosteson, T. D., Wargovich, M. J., Sandler, R., Mandel, J., Bond, J., Haile, R., Summers, R., van Stolk, R., Rothstein, R., and Weiss, J. Calcium supplementation and rectal mucosal proliferation: a randomized controlled trial. *J. Natl. Cancer Inst.*, 87: 1303-1307, 1995.
- 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.
- Alberts, D. S., Einspahr, J., Ritenbaugh, C., Aickin, M., Rees-McGee, S., Atwood, J., Emerson, S., Mason-Liddil, N., Bettinger, L., Patel, J., Bellapravala, S., Ramanujam, P. S., Phelps, J., and Clark, L. The effect of wheat bran fiber and calcium supplementation on rectal mucosal proliferation rates in patients with resected adenomatous colorectal polyps. *Cancer Epidemiol. Biomark. Prev.*, 6: 161-169, 1997.
- Schatzkin, A., Lanza, E., Freedman, L. S., Tangrea, J., Cooper, M. R., Marshall, J. R., Murphy, P. A., Selby, J. V., Shike, M., Schade, R. R., Burt, R. W., Kikendall, W., and Cahill, J. for the PPT Study Group. The polyp prevention trial I: rationale, design, recruitment, and baseline participant characteristics. *Cancer Epidemiol. Biomark. Prev.*, 5: 375-383, 1996.
- Lanza, E., Schatzkin, A., Ballard-Barbash, R., Corle, D., Clifford, C., Paskett, E., Hayes, D., Boté, E., Caan, B., Shike, M., Weissfeld, J., Slattery, M., Mateski, D., and Daston, C. for the PPT Study Group. The polyp prevention trial II: dietary intervention program and participant baseline dietary characteristics. *Cancer Epidemiol. Biomark. Prev.*, 5: 385-392, 1996.
- Searle, S. R., Casella, G., and McCulloch, C. E. *Variance Components*, pp. 2-3. New York: John Wiley & Sons, Inc., 1992.
- Lipkin, M., Blattner, W. E., Fraumeni, J. F., Jr., Lynch, H. T., Deschner, E., and Winawer, S. Tritiated thymidine (ϕ_t , ϕ_b) labeling distribution as a marker for hereditary predisposition to colon cancer. *Cancer Res.*, 43: 1899-1904, 1983.
- Grambsch, P., Louis, T. A., Bostick, R. M., Grandits, G. A., Fosdick, L., Darif, M., and Potter, J. D. Statistical analysis of proliferative index data in clinical trials. *Stat. Med.*, 13: 1619-1634, 1994.
- SAS Institute, Inc. The mixed procedure. SAS Technical Report P-229: SAS/STAT Software Changes and Enhancements, pp. 287-366. Cary, NC: SAS Institute, 1992.
- 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.
- Lyles, C. M., Sandler, R. S., Keku, T. O., Kupper, L. I., 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. Biomark. Prev.*, 3: 597-605, 1994.
- Macrae, F. A., Kiliass, D., Sharpe, K., Hughes, N., Young, G. P., MacLennan, R., and the Australian Polyp Prevention Project Investigators. Rectal epithelial cell proliferation: comparison of errors of measurement with inter-subject variance. *J. Cell. Biochem. Suppl.*, 19: 84-90, 1994.
- 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 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. Biomark. Prev.*, 4: 359-366, 1995.
- Murray, S. C., Sandler, R. S., Keku, T. O., Lyles, C. M., Millikan, R. C., Bangdiwala, S. I., Kupper, L. I., Jiang, W., and Ulshen, M. H. Comparison of rectal mucosal proliferation measured by proliferating cell nuclear antigen (PCNA) immunohistochemistry and whole crypt dissection. *Cancer Epidemiol. Biomark. Prev.*, 4: 715-720, 1995.
- Tosteson, T. D., Karagas, M. R., Rothstein, R., Ahnen, D. J., and Greenberg, E. R. Reliability of whole crypt mitotic count as a measure of cellular proliferation in rectal biopsies. *Cancer Epidemiol. Biomark. Prev.*, 5: 437-439, 1996.
- Boone, C. W., and Kelloff, G. K. (eds). Quantitative pathology in chemoprevention trials: standardization and quality control of surrogate endpoint biomarker assays for colon, breast, and prostate. *J. Cell. Biochem. Suppl.*, 19: 1-293, 1994.
- Lipkin, M., and Newmark, H. Development of clinical chemoprevention trials (Editorial). *J. Natl. Cancer Inst.*, 87: 1275-1277, 1995.
- 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. Biomark. Prev.*, 4: 57-61, 1995.
- Bostick, R. M., Fosdick, L., Lillemoe, T. J., Overn, P., Wood, J. R., Grambsch, P., Elmer, P., and Potter, J. D. Methodologic findings and considerations in measuring colorectal epithelial cell proliferation in humans. *Cancer Epidemiol. Biomark. Prev.*, 6: 931-942, 1997.
- Marra, G., Anti, M., Percesepe, A., Armelao, F., Ficarella, R., Coco, C., Rinelli, A., Vecchio, F. M., D'Arcangelo, E. Circadian variations of epithelial cell proliferation in human rectal crypts. *Gastroenterology*, 106: 982-987, 1994.
- Self, S. G., and Liang, K. Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.*, 82: 605-610, 1987.
- Morgan, E., and Gumpertz, M. Random effects models: testing whether variance components are zero. *Proc. Am. Stat. Assoc. Biom. Sec.*, 118-126, 1997.
- Pantula, S. G. Discussion on "likelihood ratio tests for variance components." *Proc. Am. Stat. Assoc. Biom. Sec.*, 127-129, 1997.