Reproducibility and Variability of the Rectal Mucosal Proliferation Index Using Proliferating Cell Nuclear Antigen Immunohistochemistry

Cynthia M. Lyles, Robert S. Sandler, Temitope O. Keku, Lawrence L. Kupper, Robert C. Millikan, Sharon C. Murray, Shrikant I. Bangdiwala, and Martin H. Ulshen


Abstract
Rectal mucosal proliferation has been shown to be increased in patients with neoplastic lesions of the large bowel and may serve as a marker of risk for colorectal malignancy. We conducted analyses to determine reliability and components of variability that might suggest optimal analysis strategies for studies of proliferation. Endoscopic pinch biopsies were obtained from 17 adult patients, labeled using proliferating cell nuclear antigen, scored using strict rules, and then rescored. Labeling index, defined as the proportion of labeled cells in a crypt, was calculated for each crypt, biopsy, subject, and group. There was excellent reproducibility. The technician was able to select previously scored crypts 95% of the time. The overall labeling index was identical on repeat. There was considerable variability in labeling index among crypts from a single biopsy and between biopsies of a single subject. Variance component estimates suggested that 20% of the variability of labeling index was due to subject, 30% due to the biopsy within a subject, and 50% due to crypts within a biopsy. There were substantial gains in statistical power by scoring two biopsies rather than one. There was less gain from further increases in biopsy number. There was little statistical advantage for counting more than 8 crypts/biopsy. Demonstrating a decrease of 25% in the mean labeling index with 90% power could require more than 100 subjects/group. We conclude that proliferating cell nuclear antigen is an extremely reproducible method to determine proliferation index. There is considerable variability among subjects, biopsies, and crypts. Analyses should use models that include random effects for subjects and biopsies to account for these sources of variability. Two to four biopsies with approximately eight scorable crypts/biopsy is the optimal number from a statistical standpoint. Large sample sizes are needed unless there are considerable differences between groups.

Introduction
Rectal mucosal proliferation index has been shown in previous studies to be elevated in both neoplastic lesions of the large bowel and in the normal mucosa of patients with colorectal neoplasia (1). As such, proliferation index has been suggested as a potential marker of risk for large bowel malignancy (2). Because this index can perhaps be manipulated with agents such as calcium (3, 4, 5), there has been hope that measurement of proliferation index might provide useful information about candidate chemopreventive agents (6). Proliferation index might be used to select individuals for interventions based on risk. Epidemiological studies might also use proliferation index to determine environmental or lifestyle exposures that correlate with elevated index values.

Classic studies of rectal mucosal proliferation used tritiated thymidine incorporation to label proliferating cells (7). In order to avoid the radiation associated with tritiated thymidine, various immunohistochemical techniques were developed (8, 9). Bromodeoxyuridine, PCNA (10, 11), and Ki67 have largely superseded tritiated thymidine because of their relative simplicity and ease of use. Studies have generally shown good correlation among all these indices, despite the fact that they label cells in different portions of the cell cycle (7, 10).

Traditionally, rectal mucosal proliferation index has been calculated by identifying well oriented crypts from mucosal biopsies, counting the number (and position) of labeled cells, and then dividing the number of labeled cells by the total number of cells for each crypt (3, 6) or hemicyrst (11). The statistical methods used in most studies are generally described without much detail. Another approach seen in the literature is to aggregate cells from all of the crypts from a patient to calculate an overall index for that patient (5, 12, 13). The position of labeled cells may also be important, with shift toward upper regions of the crypt in proliferative epithelium (14). In order to describe the positional differences in labeling, crypts are often divided into 3-5 segments from muscularis mucosa to lumen, and separate indices are calculated for each crypt segment (1, 6, 15). Preliminary work in our laboratory demonstrated con-

1 The abbreviations used are: PCNA, proliferating cell nuclear antigen; LI, labeling index.
siderable variation in the proliferation index among the crypts from biopsies of individual patients. This observed variability suggested alternate computational strategies to determine proliferation index and alternate strategies for analyzing proliferation index data in observational or intervention studies. The purpose of this paper is to evaluate the reproducibility of the proliferation index using PCNA labeling, to describe the sources of variation in proliferation, and to provide sample size nomograms for investigators contemplating large-scale, two-group comparison studies.

Materials and Methods

Biopsies were obtained from consenting patients who were referred for a clinically indicated colonoscopy. Six rectal biopsies were taken 8–12 cm from the anal verge at the start of the procedure using standard endoscopy forceps. The open cups of the forceps measure approximately 7 mm and the biopsies generally have a diameter of 2–4 mm. Each biopsy was transferred to bibulous paper (Fisher Scientific, Raleigh, NC), with special care taken not to stretch or tear the specimen. Four specimens were placed in Steinberg's modified Eagle's medium for PCNA. All specimens were processed within 30 min.

The specimens for PCNA were fixed in 70% ethanol before processing. Paraffin blocks were prepared according to routine histological procedures and care was taken not to exceed 60°C. Five sections were placed on glass slides precoated with poly-L-lysine (Sigma Chemical Co., St. Louis, MO). The sections were 5 μm thick and were taken at least 50 μm apart so that each section would contain different crypts. The sections were dewaxed and rehydrated through graded ethanol concentrations. Slides were incubated for 12–20 h at 4°C using PC10 antibody (DAKO Corp., Carpinteria, CA) at 1:100 in phosphate-buffered saline/bovine serum albumin. Detection of PCNA was performed using the Biogenex SterAvigen super sensitive kit for alkaline phosphatase (Biogenex, San Ramon, CA). The kit was used according to the manufacturer's instructions. The slides were counterstained with Mayer's hematoxylin and mounted with Eukitt (Calibrated Instruments, Inc., Hawthorne, NY). A didymium filter was added to the microscope to provide more intense staining and make labeled cells more easily recognizable. Control slides in each batch of staining included rat testis and human colon tissue from a cancer patient.

A single microscopist performed all of the scoring. Only well oriented crypts were scored. Crypts were selected for scoring if the base of the crypt touched the muscularis mucosa. If the muscularis was missing, a crypt was acceptable if its height was uniform with other crypts. For each study subject, at least 8 and at most 12 crypts were scored from each of 2 biopsies. Each of the four biopsies were screened under low power, and the two slides that had the best orientation were selected for scoring (i.e., slides with the highest number of potentially scorable crypts). Selection was based on orientation and not on the amount of staining. Beginning with the first scorable crypt in the first section, successive crypts and successive levels were counted, if necessary, until at least 8 crypts had been scored. If there were fewer than 8 scorable crypts in a biopsy, the next biopsy was selected. If there were fewer than 8 scorable crypts in 2 separate biopsies, blocks were recut and rescored. For each crypt, the total number of cells in the crypt, the center cell, and the ordinal number of stained nuclei were recorded. Only deeply stained nuclei were counted.

There are several different labeling indices that can be calculated. The ones used in our analyses are as follows. The crypt LI is defined as the proportion of labeled cells in the crypt. The biopsy mean labeling index is the sum of the labeling indices for all crypts within a given biopsy divided by the number of crypts counted. The subject mean labeling index is the sum of the labeling indices for all crypts within the subject divided by the total number of crypts. The overall mean labeling index is the sum of the labeling indices for all crypts in all subjects divided by the total number of crypts for all subjects. Each of these indices was calculated for the original and replicate samples.

Reproducibility. For reliability determinations, a random sample of 20 previously scored patients was selected from the first 60 enrolled. For 3 of the patients the specimens were inadequate, so the analyses are limited to 17 patients. Tape was placed over the label of each slide to conceal the slide information. The slides were then rescoped. The crypts were not marked so that the technician had to once again select the optimal biopsy, in terms of orientation, and then determine which crypts were scorable on that biopsy. The reliability sample therefore assesses the ability of the technician to select the same biopsy, select the same scorable crypts from a given biopsy, and calculate the same labeling index (based on numbers of labeled and unlabeled cells) for a given crypt on two different occasions.

All biopsy, slide, level, and crypt numbers were compared between the original and replicate measures to identify the number of crypts selected for scoring on both occasions. A joint 95% Bonferroni confidence region was calculated for the intercept and slope parameters of the linear regression of the replicate subject mean labeling indices on the original subject mean labeling indices to evaluate the extent of agreement between the two sets of measurements.

Variability. The variability of labeling index data can be attributed to various sources: (a) variation due to different subjects; (b) variation due to different biopsies within the same subject; and (c) variation due to different crypts within the same biopsy. Estimates of the variance components associated with these sources were obtained using restricted maximum likelihood estimation in SAS (SAS Institute, Inc.). The estimates were calculated under two different model assumptions, with and without a biopsy effect, for the original data (312 originally scored crypts). The estimates were calculated using the percentage labeled (100 × LI) as the response and also using a natural log transformation of the labeling index [Ln(100 × LI)] as the response, where Ln = natural log. The natural log transformation was performed to have the data better approximate a Gaussian or normal distribution. Using the variance component estimates, the intraclass correlation coefficients and the proportion of the total variability due to each effect were also estimated. Appendix A describes the variance components and their estimators under both model assumptions.

Because the biopsy-to-biopsy variability is often ignored, the variance component estimators are calculated under two model assumptions ("biopsy effect" and "no biopsy effect"), assuming balanced data. If the biopsy effect model holds, the no biopsy effect model is considered an underspecified model. We derive the bias (or overestima-
tion) of each variance component estimator when the underspecified model is assumed. The magnitude of the bias for each variance component estimator is evaluated by calculating the relative bias (or percentage of overestimation) when assuming the underspecified model. The formulas for the overestimation are derived in Appendix B. For various sample sizes, the relative bias in the between-subject variance component estimator is estimated for the transformed labeling index (natural log transformation) by substituting the variance component estimates from the biopsy effect model.

Sample Size Calculations. In order to provide design recommendations for future studies, minimum sample sizes necessary to attain 90% power for detecting specified group differences were calculated for various sample size combinations of subjects, biopsies, and crypts, and for various group differences. Prior to the calculations, the normality of the original labeling index data and the transformed data were assessed using a simplified version of the graphical procedure proposed by Lange and Ryan (16). The simplification involved using conservative bandwidths at ± k SD, where k = 1/2, 3/4, and 1, instead of the adjusted bandwidths described in their paper. The sample size calculations used the natural log transformation of the labeling index since the transformed data seemed to fit the normality assumption best. The sample size calculations are based on a mixed effects model (described in Appendix C), where the transformed labeling index data is modeled as the sum of an overall fixed mean, a random subject effect, a random biopsy effect, and a random error term.

The sample size calculations were based on an α-level of 0.05, using a two-sided test, the group difference of interest, the three variance components, the number of subjects/group (a), the number of biopsies/subject (b), the number of crypts/biopsy (c), and the degrees of freedom of 2 × (a-1). The formula is given in Appendix C. A population group difference of 0.2877 on the log scale, corresponding to a 25% decrease (or 33% increase) in the group mean LI, was used. The three variance component estimates were obtained from the original data using the natural log transformation of the LI.

Results

Reproducibility. Table 1 presents the descriptive statistics for the immunohistochemistry data. The original data had a total of 312 crypts scored. Three subjects had no usable tissue. For the remaining 17 subjects (one of whom had only one usable biopsy), the average number of crypts scored/subject was 18.35 (3.50, SD). The overall mean LI was 8.35% (6.41, SD). The average number of labeled cells/crypt was 8.52 (6.43, SD), and the average total cells counted/crypt was 103.48 (13.66, SD).

When the data were rescoring, a total of 313 crypts were counted. Of the 40 slides selected and scored originally, 2 slides were not scored in the replicate data. These 2 slides, which consisted of 16 scored crypts, were replaced with 2 other slides consisting of 17 scored crypts. All 296 crypts of the remaining 38 slides that were scored in the original data were successfully selected and scored in the replicate data. This means that the 296 crypts had the same biopsy, slide, level, and crypt number in both the original and replicate data. Thus, 94.9% (296 of 312) of the originally scored crypts were selected and scored on the second occasion.

Table 1 Descriptive statistics for original and replicate immunohistochemistry data

<table>
<thead>
<tr>
<th></th>
<th>Original data</th>
<th>Original data</th>
<th>Replicate data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(based on protocol)</td>
<td>(scoring same crypts as replicate data)</td>
<td></td>
</tr>
<tr>
<td>No. of biopsies scored</td>
<td>33</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>No. of crypts scored</td>
<td>312</td>
<td>313</td>
<td>313</td>
</tr>
<tr>
<td>Crypts scored per biopsy</td>
<td>9.42 (1.85)²</td>
<td>9.48 (1.86)</td>
<td>9.48 (1.86)</td>
</tr>
<tr>
<td>Crypts scored per subject</td>
<td>18.35 (3.50)</td>
<td>18.41 (3.48)</td>
<td>18.41 (3.48)</td>
</tr>
<tr>
<td>Labeling index (%)</td>
<td>8.35 (6.41)</td>
<td>8.37 (6.34)</td>
<td>8.37 (6.33)</td>
</tr>
<tr>
<td>Labeled cell counts</td>
<td>8.52 (6.43)</td>
<td>8.54 (6.36)</td>
<td>8.54 (6.36)</td>
</tr>
<tr>
<td>Total cell counts</td>
<td>103.48 (13.66)</td>
<td>103.49 (13.85)</td>
<td>103.50 (13.81)</td>
</tr>
</tbody>
</table>

²Three of the 20 subjects had unusable tissue. (Mean (sd).)

Having proven that the technician was extremely reliable at selecting biopsies and crypts to score, we next asked whether the same labeling index would be calculated for a given crypt on two separate occasions. These comparisons were made based on the common 313 crypts that were scored in both the original and replicate data sets. Table 1 presents the descriptive statistics for both the original and replicate data. The original labeling index data are displayed in Fig. 2a. Nearly all of the points in both plots lie on the 45° line going through the origin. It is clear from these figures that the technician had essentially perfect reproducibility when measuring the labeling index for a given crypt on two different occasions.

Variability. The various sources of variability in the original labeling index data are displayed in Fig. 2. a and b. Fig. 2a shows the variability due to the subject and biopsy. The overall mean labeling index is 8.35%. The subject mean labeling index ranges from about 2 to 14%. The biopsy variability within a subject is illustrated by the vertical dashed lines which connect the biopsy mean labeling indices of each subject. Fig. 2b displays the variability due to differences between crypts, both within and between biopsies. These figures demonstrate that there is considerable variation in the labeling indices between biopsies as well as within biopsies from a single subject and variation between different subjects.

The variance component estimates and estimates of their SEs are presented in Table 2. Using the untransformed data, an estimated 19.8% of the variability is due to subject differences, 30.5% of the variability is due to biopsy differences within subjects, and 49.8% of the variability is due to crypt differences within biopsies. The intraclass correlation coefficient for measures within a biopsy is estimated as 0.5204. The intraclass correlation coefficient for measures across biopsies of a subject is estimated as 0.1977. This indicates that the correlation among crypts is substantially higher when they are from the same biopsy as compared to...
Rectal Mucosal Proliferation Index using PCNA

**Fig. 1.** (a) Reproducibility of the mean labeling index percentage of the subject from repeated counts of identical crypts. Mean labeling index percentage of the subject is calculated as the mean of all labeling indices from a subject. Seventeen subjects and a total of 313 crypts were scored on each occasion. (b) Reproducibility of the crypt labeling index percentage from repeated counts of identical crypts; 313 crypts were scored on each occasion.

**Fig. 2.** (a) LI variability due to subject and biopsy differences. The overall mean LI is 8.35%. Vertical dashed lines connecting the biopsy mean labeling indices of each subject, biopsy variability within a subject. (b) Labeling index variability due to crypt differences. • and ○, crypt-labeling indices from the two biopsies within a subject.

Different biopsies of the same subject. In contrast, assuming no biopsy effect changes the composition of the variability. The variability due to subject differences is estimated as 33.4% of the total variability. Assuming no biopsy effect, it is also estimated that 66.6% of the total variability is due to crypt differences within a subject. When the data were transformed using the natural log, the figures were quite similar, with slightly higher (38.0 versus 30.5%) variability due to biopsy differences within subjects. The variance components, using the transformed data, were estimated as 0.2082, 0.4033, and 0.4493 for the between-subject, between-biopsy (within a subject), and between-crypt (within a biopsy from a subject) components, respectively.

Both the between-subject and the between-crypt variance component estimators derived using the incorrect no biopsy effect model are positively biased. The formulas for the bias are presented in Appendix B. If the between-biopsy variance component is truly equal to 0, then neither estimator is biased and the no biopsy effect model is the correct model. Table 2 shows that the between-biopsy variance component is estimated to account for 30% of the total variability (untransformed data, Model I). Since the true between-biopsy variance component is not likely to be 0, the variance component estimates for Model II in Table 2 are both overestimating the true variance components.
We estimated the relative bias (or percentage of overestimation) in the between-subject variance component estimator when assuming the underspecified no biopsy effect model. Fig. 3 presents the relative bias estimate for a range of biopsy and crypt numbers. As the number of crypts/biopsy increases and/or the number of biopsies/subject decreases, the relative bias increases. With 2 biopsies/subject and 8 crypts/biopsy, the between-subject variance component is overestimated by approximately 90% if the no biopsy effect model is assumed.

**Table 2** Variance component estimates of original labeling index data* (312 crypts)

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Assuming biopsy effect</td>
<td>Assuming no biopsy effect</td>
</tr>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
</tr>
<tr>
<td>Untransformed data (100 \cdot LI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total variability</td>
<td>41.9968</td>
<td>(8.3049)</td>
</tr>
<tr>
<td>Variability due to subjects</td>
<td>8.3045</td>
<td>(6.1630)</td>
</tr>
<tr>
<td>Variability due to biopsies within subjects</td>
<td>12.7954</td>
<td>(5.2780)</td>
</tr>
<tr>
<td>Variability due to crypts within biopsies</td>
<td>20.8969</td>
<td>(1.7698)</td>
</tr>
<tr>
<td>Correlation among crypts within biopsies</td>
<td>0.5024</td>
<td></td>
</tr>
<tr>
<td>Correlation among crypts across biopsies of a subject</td>
<td>0.1977</td>
<td></td>
</tr>
<tr>
<td>Transformed data (\ln(100 \cdot LI + 1/2))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total variability</td>
<td>1.0608</td>
<td>(0.2384)</td>
</tr>
<tr>
<td>Variability due to subjects</td>
<td>0.2082</td>
<td>(0.1736)</td>
</tr>
<tr>
<td>Variability due to biopsies within subjects</td>
<td>0.4033</td>
<td>(0.1589)</td>
</tr>
<tr>
<td>Variability due to crypts within biopsies</td>
<td>0.4493</td>
<td>(0.0381)</td>
</tr>
<tr>
<td>Correlation among crypts within biopsies</td>
<td>0.5764</td>
<td></td>
</tr>
<tr>
<td>Correlation among crypts across biopsies of a subject</td>
<td>0.1962</td>
<td></td>
</tr>
</tbody>
</table>

* Estimates are restricted maximum likelihood estimates obtained from the mixed procedure in SAS.

**Sample Size Calculations.** Fig. 4 displays the minimum sample size combinations necessary for attaining 90% power (using a two-sided test with an $\alpha$-level of 0.05) to detect a difference (for a log-transformed response) corresponding to a 25% decrease in the group mean labeling index. If only 1 biopsy was to be scored for each subject, a study with 8 crypts scored/biopsy would need at least 171 subjects/group to detect a 25% decrease in the group mean labeling index at 90% power. However, if 2 biopsies were to be scored for each subject, then a study with 8 crypts scored/biopsy would only need a minimum of 113 subjects/
group to detect the same decrease at 90% power. When the number of biopsies/subject is increased from one to two, the minimum number of subjects/group drops dramatically. When the number of biopsies/subject is increased from 2 to 3 to 4, the reduction in the number of subjects/group is less dramatic. The number of crypts/biopsy affects the minimum number of subjects/group the most when increasing from 1 to 8 crypts/biopsy. There is very little statistical advantage for counting more than 10–12 crypts/biopsy.

We performed a sensitivity analysis to see whether moderate changes in any of the variance component estimates had any effect on the sample size calculations. Eight different scenarios were compared. Two involved increasing and decreasing the total variance estimate by 20%, while maintaining the same proportions for each individual variance component. The remaining 6 scenarios fixed the total variance but altered the proportions of the components. This involved either increasing or decreasing one of the components by 20% while equally adjusting the other two components to maintain the original total variance. The overall sample size trends were unchanged under the different scenarios. There was substantial gain in power when increasing from 1 to 2 biopsies/subject but only moderate gains in power when increasing from 2 to 3 biopsies/subject. Also, there was essentially no benefit to scoring more than 8–12 crypts/biopsy. Under various scenarios the sample size generally changed by fewer than 10 subjects/group. Altering the total variance by 20% had the largest effect, changing the sample size between 15 and 35 subjects/group.

**Discussion**

In order to be useful as a measure of risk for malignancy or as an intermediate marker in intervention studies, the proliferation index must have high reliability (repeatability). Unless the histology technician can obtain the same results by scoring the same biopsy on separate occasions, then there is no value in the method. Similarly, different technicians must obtain the same results in situations where more than one technician is involved in scoring. The score ultimately depends on the proportions of the labeled cells. Cell labeling is not necessarily a yes/no decision. When using PCNA immunostaining, there are gradations of labeling differing by intensity and distribution of staining that indicate the various cell cycle phases (17). In our studies the technician must decide which cells are strongly labeled and do so in a consistent fashion. This takes training, experience, and meticulous attention to detail. We have made several modifications that help us to identify stained cells. For maximum sensitivity we use an alkaline phosphatase label rather than the horseradish peroxidase label more commonly used. We have added a dihydroxy filter to our microscope. The use of a super sensitive immunodetection system and the dihydroxy filter gives us more intense and more easily recognizable labeled cells. Scoring a slide involves more than simply deciding which cells are labeled and which are not labeled. If there are multiple biopsies, the technician must first decide which biopsy to score, then select which crypts are adequate, and only then decide which cells are appropriately labeled. The data presented in this paper show that our technician, following our protocol, was able to carry out all of these tasks with excellent reproducibility.

Over a decade ago, Deschner and Maskens (14) published data that demonstrated considerable crypt-to-crypt variation in labeling indices from human subjects and animals. Our data show that there is also considerable variability among biopsies from individual study subjects (Fig. 2, a and b). Previous studies have considered the crypt as the unit of observation but they have failed to account for the biopsy from which the crypts were derived (11, 18). Both the subject and the biopsy are significant sources of variability that must be considered in the analysis of immunohistochemistry data. Fig. 3 demonstrates that failure to account for the between-biopsy variance component results in a positively biased between-subject variance component estimate. The overestimation can be as large as 80% if two biopsies are scored per subject and as large as 190% if only one biopsy is scored per subject. Table 2 also shows that the between-subject and between-crypts variance components are overestimated if the between-biopsy variance component is ignored. We would therefore recommend that investigators who use proliferation index in research studies analyze the data with models that account for subject and biopsy effects.

There are additional concerns with the underlying models assumed when analyzing proliferation data. Previous authors have sometimes aggregated all of the cells from all crypts and biopsies of a patient to determine an overall labeling index for each patient (5, 12, 13). The test procedures that have been used to analyze these data rely on models that assume all of the cells within a crypt are independent. This is not likely to be the case. In addition, analyses of this type may not provide any sort of weighting for the unequal contribution of cells. Patients with more crypts or larger crypts (and thus more cells) contribute more information than patients with fewer or smaller crypts. Other authors (3, 4) have calculated a labeling index for each crypt and a subject mean labeling index as a mean of all crypt-labeling indices within the subject. The test procedures used, however, to assess a difference in the mean labeling index between groups, when using the subject mean labeling index as the response, require the assumption that all subjects mean labeling indices have equal variance. This homogeneity assumption is violated if the number of crypts within biopsies or biopsies within subjects varies between subjects. The appropriate statistical model should account for the unbalanced nature of the data.

Other important statistical assumptions typically placed on the data are the normality assumption and the homogeneity of variance assumption. As shown in Fig. 1b, proliferation data are not normally distributed but are extremely skewed to the right. The labeling index ranges between 0 and 41%, with the majority of the labeling indices less than 15%. The natural log transformation of the labeling index seems to improve the normality of the data. The homogeneity of variance was not directly checked in our data. However, most normalizing transformations also improve potential heterogeneity of variance in the data (19). Assessing the distributional assumptions placed on the data and using a transformation of the data if needed is recommended prior to analysis of proliferation data.

The sample size calculations are interesting in several respects. First, they provide some guidance about the minimal number of biopsies to obtain and crypts to count. There is substantial gain in power that is achieved by increasing the number of biopsies from one to two. Further increases in the number of biopsies are associated with more modest gains in power. The power curves also demonstrate that there are no real gains in power by counting more than eight crypts from a given biopsy. Moreover,
simply counting more crypts is not equal to the gains in power that are achieved by increasing the number of biopsies. Determination of proliferation index is extremely labor intensive. Although counting labeled cells is tedious and exacting work, preparing, sectioning, and orienting biopsies is even more labor intensive. Thus, the major laboratory “cost” is the time to process and prepare the individual slides (biopsies). In order to minimize the cost, the investigator would endeavor to minimize the total number of biopsies needed to be scored per group, while maintaining a 90% power for detection of group differences. According to our calculations, when scoring 8 crypts/biopsy, the power would be identical with one biopsy for each of 171 subjects (171 biopsies), two biopsies on 113 subjects (226 total biopsies), or three biopsies on 93 subjects (279 biopsies). Although the number of subjects needed drops as more biopsies are taken, the total number of biopsies increases. It is generally more difficult to recruit additional study subjects than it is to obtain more biopsies from subjects who have already been recruited. Based on recruitment and laboratory cost/effort issues, we would recommend that investigators score three biopsies on 93 subjects (279 biopsies) as this is the labor intensive cost that are achieved by increasing the number of biopsies.

It is also apparent from the power curves that large numbers of subjects are necessary to demonstrate a 25% decrease between groups. Investigators who are planning intervention or observational studies using proliferation index as an outcome variable need to be aware that over 100 subjects are needed in each group to demonstrate differences of this magnitude. Obviously, differences between groups that are larger will require fewer subjects. For example, to detect a 50% decrease in the mean labeling index at 90% power with 2 biopsies/subject and 8 crypts/biopsy, a study would require 21 subjects/group. Some previous reports have demonstrated an effect of calcium on proliferation with small numbers of subjects (3, 4, 5, 18), although other studies have failed to confirm these results (13, 20, 21).

Previous authors have also performed sample size calculations for detecting group differences in proliferation (21, 22). The sample size recommendations in each of these reports vary somewhat. We, as well as Freedman and Schatzkin (22), assume that the labeling index data are log normal. Bostick et al. (21) do not. The assumed model to describe the labeling index data differs in each report resulting in differences in sample size estimates. For example, we assumed a mixed model of the crypt-transformed labeling index, which included nested random effects for subject, biopsy, and crypt to account for the corresponding sources of variation (see Appendix C). The estimated variance of the group mean labeling index used in the sample size calculations, which depends on the characteristics of the sample population from which the variance estimate is obtained and the labeling methods used to measure proliferation, also varies. Lastly, the prespecified parameters in the sample size calculations, such as the power, $\alpha$-level, 1 or 2-sided test procedure, and the difference in group mean labeling index to detect are different in each of these reports. Changing any one of these parameters can significantly affect the sample size recommendations.

There are several labeling methods that have been used to determine rectal proliferation index. The results presented in this paper are specific for PCNA. We have done previous power analyses of data obtained using tritiated thymidine and found similar results on the number of crypts to be scored but smaller numbers of subjects needed in each group. Nonetheless, there is the potential for variable results depending on labeling method and the laboratory. However, the conclusions for design recommendations based on the shape of the power curves are likely to be similar.

In summary, we have shown that the method of PCNA immunohistochemistry determination of rectal mucosal proliferation is highly reproducible. There is considerable variation from crypt-to-crypt and from biopsy-to-biopsy for the same individual. Taking this variation into account during analysis of proliferation data will produce more valid statistical estimates and inferences.

Acknowledgments
We thank Wuhan Jiang for technical assistance.

Appendices

Appendix A

Let $Y_{ijkl}$ be the set of labeling indices where $Y_{ijkl}$ is the labeling index for the $i^{th}$ crypt ($i = 1, 2, \ldots, c$), within the $k^{th}$ biopsy ($k = 1, 2, \ldots, b$), from the $j^{th}$ subject ($j = 1, 2, \ldots, a$), in the $l^{th}$ group ($l = 1, 2$).

Biopsy Effect Model: Two-Way Nested Classification Mixed Model with Balanced Data. Under the model assuming a random subject effect and a random biopsy effect, the total variability $(\sigma^2_t)$ of a labeling index can be decomposed as $\sigma^2_t = \sigma^2_s + \sigma^2_b + \sigma^2_c$, where $\sigma^2_s$ is the variability between subjects, $\sigma^2_b$ is the variability between biopsies within a subject, and $\sigma^2_c$ is the variability between crypts within a biopsy.

The corresponding analysis of variance table (23) is created by calculating the sums of squares and mean squares, as if the model is a fixed effects model, and then deriving the expected values of the mean squares under the assumed mixed model (see Table A1).

The analysis of variance method of estimating the variance components leads to the following variance component estimates.

\[ \hat{\sigma}^2_s = \text{MSE}_{\text{BC}}; \quad \hat{\sigma}^2_b = \frac{\text{MSE}_{\text{BC}} \cdot \text{MSE}_{\text{A}}}{c}; \quad \hat{\sigma}^2_c = \frac{\text{MSE}_{\text{A}} - \text{MSE}_{\text{BC}}}{b} \]

No Biopsy Effect Model: One-Way Nested Classification Mixed Model with Balanced Data. Under the model assuming only a random subject effect, the total variability $(\sigma^2_t)$ of a labeling index can be decomposed as $\sigma^2_t = \sigma^2_s + \sigma^2_c$, where $\sigma^2_s$ is the variability between subjects and $\sigma^2_c$ is the variability between crypts within a subject.

The corresponding analysis of variance table (23) is created in the same manner as described above. The expected values of the mean squares are derived assuming the no biopsy effect model (see Table A2).

The analysis of variance method of estimating the variance components leads to the following variance component estimates.

\[ \hat{\sigma}^2_s = \text{MSE}_{\text{BC}}; \quad \hat{\sigma}^2_c = \frac{\text{MSE}_{\text{BC}} \cdot \text{MSE}_{\text{A}}}{b} \]

Appendix B

Bias in Variance Component Estimators Using Incorrect No Biopsy Effect Model. If the biopsy effect model is true and the No Biopsy Effect Model is assumed, the resulting variance component estimators are biased. $\text{Bias}(\hat{\sigma}^2_c) = (E(\hat{\sigma}^2_c) - \sigma^2_c)$ is calculated as the difference between the expectation, under the true model, of the variance component estimator, derived from the underspecified model, and the variance component parameter.

The variance component estimators derived from the underspecified model (no biopsy effect model in Appendix A) can be rewritten as:

\[ \hat{\sigma}^2_c = \frac{\text{MSE}_{\text{BC}} \cdot \text{MSE}_{\text{A}}}{b} \cdot \frac{1}{c} \cdot \frac{1}{\left[ \sum_{i=1}^{c} \sum_{j=1}^{a} \sum_{l=1}^{3} (Y_{ijkl} - \bar{Y}_i)^2 \right] \cdot \left[ \sum_{j=1}^{a} \sum_{l=1}^{3} (\bar{Y}_i - \bar{Y}_j)^2 \right]} \]

and

\[ \hat{\sigma}^2_c = \frac{\text{MSE}_{\text{BC}} \cdot \text{MSE}_{\text{A}}}{b} \cdot \frac{1}{c} \cdot \frac{1}{\left[ \sum_{j=1}^{a} \sum_{l=1}^{3} (Y_{ijkl} - \bar{Y}_i)^2 \right] \cdot \left[ \sum_{j=1}^{a} \sum_{l=1}^{3} (\bar{Y}_i - \bar{Y}_j)^2 \right]} \]

\[ \hat{\sigma}^2_c = \frac{\text{MSE}_{\text{BC}} \cdot \text{MSE}_{\text{A}}}{b} \cdot \frac{1}{c} \cdot \frac{1}{\left[ \sum_{j=1}^{a} \sum_{l=1}^{3} (Y_{ijkl} - \bar{Y}_i)^2 \right] \cdot \left[ \sum_{j=1}^{a} \sum_{l=1}^{3} (\bar{Y}_i - \bar{Y}_j)^2 \right]} \]
Table A1

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>E( MS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>1</td>
<td>SSA = abc (\sum_{i=1}^{2} (\hat{Y}_i - \bar{Y})^2)</td>
<td>SSA (\frac{t}{1})</td>
<td>(\sigma_c^2 + \sigma_2 + bc\sigma_s^2 + abc \sum_{i=1}^{2} \mu_i^2)</td>
</tr>
<tr>
<td>Subject</td>
<td>2(a-1)</td>
<td>SSB 1A = bc (\sum_{i=1}^{2} \sum_{j=1}^{j} (\hat{Y}<em>{ij} - \bar{Y}</em>{ij})^2)</td>
<td>SSB 1A (\frac{2a(1-1)}{2})</td>
<td>(\sigma_c^2 + \sigma_2 + bc\sigma_s^2)</td>
</tr>
<tr>
<td>Biopsy</td>
<td>2(a-b-1)</td>
<td>SSC 1B = c (\sum_{i=1}^{2} \sum_{j=1}^{j} (\hat{Y}_{i} - \bar{Y}_i)^2)</td>
<td>SSC 1B (\frac{2a(b-1)}{2})</td>
<td>(\sigma_2 + \sigma_s^2)</td>
</tr>
<tr>
<td>Error</td>
<td>2(a-bc-1)</td>
<td>SSE = (\sum_{i=1}^{2} \sum_{j=1}^{j} \sum_{k=1}^{k} (Y_{ij} - \bar{Y}_{ij})^2)</td>
<td>SSE (\frac{2a(b-1)}{2})</td>
<td>(\sigma_c^2)</td>
</tr>
<tr>
<td>Total</td>
<td>2abc-1</td>
<td>SST = (\sum_{i=1}^{2} \sum_{j=1}^{j} \sum_{k=1}^{k} (Y_{ij} - \bar{Y}_{ij})^2)</td>
<td>SST (\frac{2abc-1}{2})</td>
<td>(\sigma_2 + \sigma_s^2)</td>
</tr>
</tbody>
</table>

Table A2

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>E( MS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>1</td>
<td>SSA = abc (\sum_{i=1}^{2} (\hat{Y}_i - \bar{Y})^2)</td>
<td>SSA (\frac{t}{1})</td>
<td>(\sigma_c^2 + bc\sigma_s^2 + abc \sum_{i=1}^{2} \mu_i^2)</td>
</tr>
<tr>
<td>Subject</td>
<td>2(a-1)</td>
<td>SSB 1A = bc (\sum_{i=1}^{2} \sum_{j=1}^{j} (\hat{Y}<em>{ij} - \bar{Y}</em>{ij})^2)</td>
<td>SSB 1A (\frac{2a(1-1)}{2})</td>
<td>(\sigma_c^2 + bc\sigma_s^2)</td>
</tr>
<tr>
<td>Error</td>
<td>2(a-bc-1)</td>
<td>SSE = (\sum_{i=1}^{2} \sum_{j=1}^{j} \sum_{k=1}^{k} (Y_{ij} - \bar{Y}_{ij})^2)</td>
<td>SSE (\frac{2a(b-1)}{2})</td>
<td>(\sigma_2 + \sigma_s^2)</td>
</tr>
<tr>
<td>Total</td>
<td>2abc-1</td>
<td>SST = (\sum_{i=1}^{2} \sum_{j=1}^{j} \sum_{k=1}^{k} (Y_{ij} - \bar{Y}_{ij})^2)</td>
<td>SST (\frac{2abc-1}{2})</td>
<td>(\sigma_2 + \sigma_s^2)</td>
</tr>
</tbody>
</table>

Now, the expectations of the above estimators under the true biopsy effect model are derived.

\[E(\sigma^2) = \frac{1}{2abc-1} \left\{ \sum_{i=1}^{2a} \sum_{j=1}^{2b} \sum_{k=1}^{2c} (Y_{ijk} - Y_{ij})^2 \right\} - E(MSE)\]

\[= \frac{1}{2abc-1} \left\{ \sum_{i=1}^{2a} \sum_{j=1}^{2b} \sum_{k=1}^{2c} (Y_{ijk} - Y_{ij})^2 \right\} - E(MSE)\]

\[= \left\{ \sigma_c^2 + c(b-1)\sigma_s^2 + 2a(b-1)(\sigma_2 + \sigma_s^2) \right\}\]

and

\[E(\sigma^2) = \frac{1}{2abc-1} \left\{ \sum_{i=1}^{2a} \sum_{j=1}^{2b} \sum_{k=1}^{2c} (Y_{ijk} - Y_{ij})^2 \right\} - E(MSE)\]

\[= \frac{1}{2abc-1} \left\{ \sum_{i=1}^{2a} \sum_{j=1}^{2b} \sum_{k=1}^{2c} (Y_{ijk} - Y_{ij})^2 \right\} - E(MSE)\]

\[= \left\{ \sigma_c^2 + c(b-1)\sigma_s^2 + 2a(b-1)(\sigma_2 + \sigma_s^2) \right\}\]

It follows that the bias of the variance component estimators under the incorrect no biopsy effect model are:

\[\text{Bias}(\sigma_c^2) = \frac{c(b-1)}{bc-1} \sigma_s^2\]

Table A2

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>E( MS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>1</td>
<td>SSA = abc (\sum_{i=1}^{2} (\hat{Y}_i - \bar{Y})^2)</td>
<td>SSA (\frac{t}{1})</td>
<td>(\sigma_c^2 + \sigma_2 + bc\sigma_s^2 + abc \sum_{i=1}^{2} \mu_i^2)</td>
</tr>
<tr>
<td>Subject</td>
<td>2(a-1)</td>
<td>SSB 1A = bc (\sum_{i=1}^{2} \sum_{j=1}^{j} (\hat{Y}<em>{ij} - \bar{Y}</em>{ij})^2)</td>
<td>SSB 1A (\frac{2a(1-1)}{2})</td>
<td>(\sigma_c^2 + \sigma_2 + bc\sigma_s^2)</td>
</tr>
<tr>
<td>Biopsy</td>
<td>2(a-b-1)</td>
<td>SSC 1B = c (\sum_{i=1}^{2} \sum_{j=1}^{j} (\hat{Y}_{i} - \bar{Y}_i)^2)</td>
<td>SSC 1B (\frac{2a(b-1)}{2})</td>
<td>(\sigma_2 + \sigma_s^2)</td>
</tr>
<tr>
<td>Error</td>
<td>2(a-bc-1)</td>
<td>SSE = (\sum_{i=1}^{2} \sum_{j=1}^{j} \sum_{k=1}^{k} (Y_{ij} - \bar{Y}_{ij})^2)</td>
<td>SSE (\frac{2a(b-1)}{2})</td>
<td>(\sigma_c^2)</td>
</tr>
<tr>
<td>Total</td>
<td>2abc-1</td>
<td>SST = (\sum_{i=1}^{2} \sum_{j=1}^{j} \sum_{k=1}^{k} (Y_{ij} - \bar{Y}_{ij})^2)</td>
<td>SST (\frac{2abc-1}{2})</td>
<td>(\sigma_2 + \sigma_s^2)</td>
</tr>
</tbody>
</table>

Appendix C

The two-way nested classification mixed model for balanced data (23) used for sample size calculations is:

\[Y_{ijk} = \mu + \alpha_{ij} + \gamma_{ij} + \epsilon_{ijk}\]

where \(i = 1, 2\) groups; \(j = 1, 2, \ldots, \) (a) subjects per group; \(k = 1, 2, \ldots, \) (b) biopsies per subject; and \(i = 1, 2, \ldots, \) (c) crypts per biopsy.
This model involves the following assumptions: (a) $\mu_i$ is the fixed effect, or mean natural log labeling index of the $i^{th}$ group; (b) $\sigma_{0i} \sim \text{N}(0, \sigma^2)$ is the random effect of the $i^{th}$ subject in the $i^{th}$ group; (c) $Y_{ik} \sim \text{N}(Y_i, \sigma^2)$ is the random effect associated with the measure from the $i^{th}$ group. 

The above assumptions imply that: (a) $E[Y_{ik}] = \mu_i$ for all $i$, $k$; (b) for $i = 1, 2$; (c) $\text{Corr}(Y_{ik}, Y_{ij}) = 0$, for all $i \neq j$, $k$, $l$; (d) $\text{Corr}(Y_{ik}, Y_{ij}) = (\sigma^2_1 + \sigma^2_2 + \sigma^2_3)$, for all $k \neq k'$, is the correlation among measures from different biopsies within a subject; and (e) $\text{Corr}(Y_{ik}, Y_{i'j}) = [(\sigma^2_1 + \sigma^2_2)(\sigma^2_3 + \sigma^2_4 + \sigma^2_5)]$, for all $i \neq i'$, is the correlation among measures within a biopsy from a subject.

The power function for detecting a specified treatment effect ($\mu_1 - \mu_2$) at the $\alpha$ level, using a two-sided test, under the proposed mixed model has the following form:

$$
\text{Power} = P\{T > t_{2a,1-\alpha/2} \mid \delta \}
$$

where $t_{2a,1-\alpha/2}$ is the critical value of the $t$ distribution with $2(a-1)$ degrees of freedom at the $1 - \alpha/2$ percentile.

$\delta = \sqrt{\frac{2 \cdot (\sigma^2_1 + \sigma^2_2/a \cdot b + \sigma^2_3/a \cdot c \cdot b \cdot c)}{2 \cdot (a \cdot b)}},
\text{is the non-centrality parameter of the appropriate non-central t distribution,}
\delta = \sqrt{\frac{2 \cdot (a \cdot b)}{2 \cdot (a \cdot b)}},
\text{is the test statistic used to test } H_0: \mu_1 = \mu_2 \text{ versus } H_1: \mu_1 \neq \mu_2, \text{where } Y_i
\text{is the mean of the natural log transformation of the labeling index and a, b, c, } \sigma^2_1, \sigma^2_2, \text{and } \sigma^2_3 \text{are described above.}
\text{The power is based on a } \alpha \text{-level of 0.05, using a two-sided test, a treatment effect size } (\mu_1 - \mu_2), \text{values for a, b, c, and estimates of the variance }
\text{components } \sigma^2_1, \sigma^2_2, \text{and } \sigma^2_3 \text{from the natural log transformation of the labeling index data. The power was calculated using the PROBF function in SAS (SAS Institute, Inc.). The sample size combinations were obtained by finding the minimum sample size combinations that achieved a power of at least 0.90.}

References:
Reproducibility and variability of the rectal mucosal proliferation index using proliferating cell nuclear antigen immunohistochemistry.

C M Lyles, R S Sandler, T O Keku, et al.