Analysis of Cell Proliferation Data

by Richard W. Morris

When estimating the labeling index is of interest, the design of experiments raises a number of methodological questions: How many cells should be scored? How big a difference in labeling index is likely to be detectable? What is the potential effect of low growth fraction on detecting a treatment effect? What are appropriate ways of expressing treatment effects on labeling index? Data from two labeling index experiments are used to shed light on these questions. The answers to all questions depend on the level of labeling index under consideration: a low frequency of labeling makes it important to count more cells, but this should not be done at the expense of using fewer animals. Detecting differences between treated and control groups when labeling index is low or when growth fraction is low is difficult, and caution must be used when expressing treatment effect as fold increase when labeling index is small.

Introduction

The empirical study of cell proliferation depends on the ability to distinguish cells engaged in the process of cell replication from those that are not. A variety of methods of cytokinetic analysis can identify cells that have entered one or more stages of the cell cycle. One of the most well-known and intuitively appealing methods is to label cells that are actively synthesizing nuclear DNA and to estimate the percentage of nuclei that exhibit label. This quantity, known as the labeling index (LI), is frequently used as a measure of response in experiments seeking to characterize the effect of treatment on cell proliferation. In this article, I discuss some practical aspects of estimating LI and testing hypotheses about treatment effect on LI. I also consider the effect of growth fraction on detecting treatment effect and alternative ways of expressing treatment effects on LI.

Estimating Labeling Index

The first question that usually comes to mind in planning an experiment to estimate LI is, How many cells should I count? The short answer is that the more cells you count, the smaller the variance of the estimate of LI. A small variance is desirable, but this answer is unacceptably vague and overlooks the possibility of variation in LI among regions within a tissue or among animals. When these sources of variation are present, as experimental data suggest they are, then counting more cells on each microscope slide will reduce the variance of LI, but only up to point, beyond which additional counting has little effect on reducing the variance of LI. The point of diminishing returns depends on LI, but in general, more cells should be counted when LI is small than when LI is large.

To see why the benefits of additional counting diminish, assume for a particular region of tissue to be sampled that a fixed fraction of cells are asynchronously cycling and that the chance of any cell in S phase becoming labeled is constant. An estimate of LI can be obtained as the number of labeled cells/number of cells scored. This fraction multiplied by 100 is defined as LI, but by itself can be interpreted as an estimate of the probability that a cell enters S phase when label is present. This estimate has binomial variance \( p(1-p)/c \), where \( p \) is the probability of label and \( c \) is the number of cells scored.

As \( c \) increases, the variance of the estimate of \( p \) decreases, yielding a more precise estimate of \( p \). If we apply this argument separately to each region of tissue in each animal, we are led to the conclusion that an increase in cells scored yields a decrease in the variance of LI. But this argument overlooks the possibility that \( p \) itself varies among regions within a tissue or among animals.

Now assume that \( p \) varies among animals and we wish to estimate the mean \( p \). For simplicity, assume that \( p \) does not vary among regions of a tissue. Then an estimate of \( p \) for an individual has two sources of variation: the variance due to sampling nuclei, noted earlier, and the variance due to sampling individuals. The total variance of \( p \) can be expressed as the sum of these two
components \( p(1-p)/c + \sigma^2 \), where \( \sigma^2 \) is the variance of \( p \) among animals. Notice that when \( \sigma^2 \) is greater than zero, that is, when \( p \) varies among individuals, the variance of an estimate of \( p \) for an individual can never be less than \( \sigma^2 \), no matter how many cells are counted.

Because interferences are sought for the population from which treatment groups are drawn, the mean LI for a group of animals rather than the LI for an individual is of interest. In this case, the variance of the estimated mean of \( p \) depends on the number of animals in the group as well as on the number of cells scored per animal. This variance is given by

\[
\frac{\sigma^2(c-1) + \bar{p}(1-\bar{p})}{\eta c},
\]

where \( \eta \) is the number of animals in the group and \( \bar{p} \) is the mean \( p \) among animals in the group. When the number of cells scored is more than 100, this expression is approximately

\[
\frac{\sigma^2}{\eta c} + \frac{\bar{p}(1-\bar{p})}{\eta c}.
\]

Figure 1 shows a contour plot of the variance of \( p \) as a function of the number of animals, \( n \), and the number of cells scored, \( c \). Darker regions of the plot correspond to larger values of variance, and lighter regions of the plot correspond to smaller values of variance. The contours indicate that the greater the number of cells counted and the greater the number of animals examined, the smaller the variance of LI. But for the data

\[\text{set from which Figure 1 was drawn, the trade-off is not even. For any given number of animals, an increase in number of cells counted reduces the variance of LI, but it only does so appreciably for the first few hundred cells counted. After a few hundred cells have been scored, little further reduction in variance is achieved by counting more cells.}\]

The data from which Figure 1 was drawn are from counts of 1000 bromodeoxyuridine (BrdU)-labeled hepatocytes per male mouse liver. The mean LI was 5.4 and its variance was 21.5. This total variance was estimated as the residual in an analysis of variance with 12 groups of 8–10 animals per group. The among-animal component of variance used in the contour plot was obtained by setting the expected total variance, given above, to the estimated value and solving for the among-animal component.

Figure 2 shows a contour plot of a different data set. In this data set, 10 groups of 9–10 animals yielded a mean BrdU label index of 1.25 with a total variance of 5.28. The contours in Figure 2 exhibit more pronounced curvature than in Figure 1. As a result, for a given number of animals, an increase in the number of cells counted will continue to yield a reduction of the variance of the mean LI for much larger numbers of cells counted than the data plotted in Figure 1 indicate.

A comparison of the contour shapes in Figures 1 and 2 suggests two generalizations. First, the number of cells counted per animal should be inversely related to the mean LI to be estimated. When LI is small, say 1% or less, more effort should be spent estimating its value in each tissue sample than when LI is much larger. It is reasonable to expect to score more cells when those exhibiting label are rare than when they are common. The implication of this conclusion for experi-
Detecting a Treatment Effect

One of the goals of cell proliferation experiments is to detect a difference, when one exists, between treated and control groups. The ability to detect a treatment effect on mean LI depends on a number of characteristics of the experiment. First, and perhaps of most interest, is the size of the difference in mean LI between control and treated groups brought about as a consequence of treatment. In general, a small change in mean LI will be more difficult to detect than a large change. In addition to the size of the effect on LI, the number of observations, the amount of variation among slides, tissues, and animals in LI, and the statistical test used all influence the chance of detecting a treatment-related effect.

One way of combining the factors that influence an experimenter's ability to detect a treatment effect is to plot the power of the statistical test. Figures 3 and 4 plot the power of the well-known t-test to detect a treatment difference in mean LI from a control value of 5 (Fig. 3) or from a control value of 1 (Fig. 4). Power curves for group sizes of 4, 8, and 12 are presented. The variances estimated from the two data sets described earlier were used to produce these plots.

Because the power of a test is defined as the probability of rejecting the hypothesis of no treatment effect when, in fact, there is a treatment effect, we expect the power of the t-test to increase as the mean LI of the treated group becomes more different from the control LI. In Figure 3, a doubling of the control mean of 5, yielding a treatment mean of 10, has probability 0.4 of being detected in groups of size 4. If group size is increased to 8 or 12, the power of the t-test increases to approximately 0.7 or 0.9, respectively. Thus, with the amount of variation seen in data with a mean LI of 5 and group sizes of 10 or more, there is a good chance of detecting a doubling in LI. On the other hand, when the mean LI of the control is small, as in Figure 4 where the control LI is 1, a 3- to 4-fold increase over the control mean is required to achieve power similar to that obtained when the control LI is larger.

The use of power curves associated with a t-test may be open to question when applied to LI data because this test assumes that data are normally distributed and that the variances of the two treatment groups are the same. Neither of these assumptions is likely to hold for LI data. The distribution of LI is generally asymmetric, with a clump of values to the left toward 0 and a long tail of values extending to the right. Moreover, the variance of LI typically increases with the mean LI. Nevertheless, the t-test is generally recognized as robust to failure of these assumptions, and it often serves as a valid statistical test even when all assumptions are not met.

To help evaluate whether the t-test is usefully applied to LI data, computer simulation comparing the power of the t-test with the power of the rank sum test
was carried out for LI data. The rank sum test does not have strong distributional assumptions like the t-test and therefore may be more appropriate for LI data.

Figure 5 shows power curves obtained for groups of size 8 when the control exhibits a mean LI of approximately 5. Two power curves are presented, one for the t-test and one for the rank sum test. A t-test based on a transformation of LI, in which the arcsine of the square root of LI/100 was used, had power virtually identical to that of the rank sum test.

The power curves in Figure 5 are very close together, suggesting that the t-test does not suffer a great loss of power due to the extrabinomial variation in this example. This quick look at several approaches to hypothesis testing suggests there is little to choose between the t-test and the rank sum test, although there may be circumstances that occur in LI data that are unfavorable to exclusive use of the t-test.

Growth Fraction

One of the difficulties in working with LI is the uncertainty in interpreting a change in LI in the presence of an unknown growth fraction. Certainly, a mitogen that produces an increase in LI of 5 would be viewed with far more interest if it were known that the growth fraction was 10% than if it were 90%. This is so because if a control LI of 1 was to be increased to 6 by treatment when the growth fraction was 10%, it would mean that 5 additional cells of every 10 cells that were actively cycling would be entering S phase as a consequence of treatment. This represents a dramatic increase in the frequency of cycling cells entering S phase. In contrast, with a growth fraction of 90%, only 5 additional cells of every 90 cells actively cycling would be recruited into S phase. Thus, a given change in LI may have different implications concerning proliferation response, depending on the growth fraction.

How does growth fraction affect the ability to detect a treatment response as reflected by a change in LI? Assume for illustration that an agent acts as a pure mitogen and its effect in a given protocol is to double the number of cells entering S phase. With a background LI of 25 and a growth fraction of 100%, the treatment response would be an LI of 50. The same cell kinetics with a growth fraction of 50% would yield control and treatment LIs of 12.5 and 25, respectively. When the growth fraction is 1%, control and treatment LIs drop to 0.25 and 0.50. Thus, as the growth fraction decreases, the absolute difference between control and treated LI decreases, and, not surprisingly, this difference becomes harder to detect.

Figure 6 plots the power of a t-test against growth fraction when the effect of treatment is to double the probability of a cell entering S phase from 0.25 to 0.5. Power was computed by assuming that binomial sampling was the sole source of variation. Three sample sizes are plotted: 100, 500, and 1000 cells. Power drops off steadily with decreasing growth fraction with samples of size 100. For larger sample sizes of 500 and 1000, power remains high until growth fraction falls below 20%, then power drops precipitously. These observations suggest that as growth fraction becomes small, treatment-induced changes in cell cycle kinetics will become harder to detect.

Expressing Treatment Effects

There are two generally accepted ways to express change in LI as a consequence of treatment. The simplest way is to list the mean and standard error of LI for each treatment in a study. An alternative is to express the LI of treated groups as a ratio: treated

![Graph](image1)

**Figure 5.** Power of the t-test and rank sum test. Power obtained from computer simulation described in text.

![Graph](image2)

**Figure 6.** Power of the t-test against growth fraction. Power was computed for binomial samples of size \( c = 100, 500, \) and 1000 cells.
LI/control LI. This latter approach is referred to as fold increase (or decrease). Fold increase is a unitless statistic that makes comparisons of response between studies straightforward. Although fold increase has appeal in its simplicity, caution should be exercised in its interpretation because the variance of this statistic is a complex function of the means and variances of control and treated groups. The variance of fold increase is rarely reported.

An approximate expression for the variance of the ratio of treated LI/control LI is given by

\[
\sigma^2_L = \frac{\sigma^2_0}{L_0^2} + \frac{\sigma^2_1}{L_1^2} + \frac{\sigma^2_0 \cdot L_1^2}{L_0^2 \cdot L_1^2}
\]

where \( L \) represents labeling index, \( \sigma^2 \) is the variance of LI and subscripts 0 and 1 refer to control and treated groups, respectively. Because the mean of the control shows up in the denominator, the behavior of the variance of fold increase when the control LI becomes small is of interest. Assume, again for simplicity, that the only source of variation in estimating the mean LI is binomial sampling. This gives a lower bound on the variance of LI. When the variance of fold increase is plotted against the control LI, the result shown in Figure 7 is obtained.

Three values of fold increase, ratios 2, 3, and 4, are plotted. As the control LI becomes small, the variance of fold increase becomes large. This observation implies that the same fold increase in two studies may have widely differing precision, depending on the value of the control group LI. Thus, when fold increase is reported, its standard error should also be reported. Failing this, the control LI should be clearly stated along with fold increase. The foregoing considerations suggest that fold increases associated with small control values of LI should be interpreted with caution because they will have large variance.

**Summary**

The preceding discussion leads to several generalizations. First, it is important to keep in mind, when designing a study to estimate LI, that the number of cells counted per slide is only one of a number of potential sources of variation in mean LI. Consideration of additional sources of variation, such as among tissues or among animals, shows that it is important not to focus solely on the individual microscope slide as the sampling unit. To do so may lead to reduced numbers of tissues or animals sampled with the possible consequence of increased variance of mean LI.

Second, some relatively simple considerations of statistical power can lead to a more clearer understanding of the amount of change in LI that is likely to be detectable, in a particular experimental system. For the limited data examined here, a doubling of LI when the control is near 5 or a quadrupling of LI when the control is near 1 appears to be detectable with a rank sum test or a t-test.

Third, caution should be exercised when expressing a change in LI as fold increase. Unless its variance is stated or the control mean is also noted, fold increase may be a potentially misleading statistic to use in characterizing the results of treatment effect because the uncertainty associated with an estimate of fold increase is unclear when this statistic is reported alone.

I thank R. R. Maronpot for providing the data used in this article.

**APPENDIX**

The computer simulation of power for the t-test and rank sum test consisted of 1000 experiments for each of 10 values for the treated group mean LI. Each group had 8 individuals. The LI values were generated in two stages, reflecting variation among individuals and variation within individuals as a consequence of binomial sampling. For each individual, a value for LI was drawn from a \( \beta \) distribution with parameters \( \alpha=1 \) and \( \beta=20 \), yielding a mean of 0.048 and SD of 0.045 for the control group. The value of LI was used to generate a binomial observation with 100 cells sampled. Treatment group LIs were generated by sequentially increasing \( \alpha \). The simulation was done using the SAS system.

**REFERENCES**

1. Gray, J. W., and Darzynkiewicz, Z., Eds. Techniques in Cell Cycle Analysis. Humana Press, Clifton, NJ, 1989, p. 407.
2. Appleton, D. R. Aspects of statistics in studies of cell proliferation I. Multistage sampling. Cell Tissue Kinet. 17: 545-548 (1984).
3. Gundersen, H. J. G., and Osterby, R. Optimizing sampling efficiency of stereological studies in biology: or Do more less well? J. Microsc. 121: 65–72, (1981).