Hiroshima Journal of Medical Sciences
Vol.34, No.1, 31~37, March , 1985
HIJM 34-5

Some Statistical Issues Pertaining to *in Vitro* Drug Testing with Human Tumor Colony Forming Assays

James A. KOZIOL

Department of Epidemiology and Statistics, Radiation Effects Research Foundation, 5-2 Hijiyama Park, Mnami-ku, Hiroshima 730, Japan

(Received November 15, 1984)

Key words: In vitro drug testing, Salmon assay

ABSTRACT

One should proceed with prudence in the interpretation and application of *in vitro* drug tests. Among the statistical considerations that should be borne in mind are the following:

1. Reproducibility of *in vitro* assays should be adequately addressed.
2. Experimental designs for *in vitro* assays need further development, and should take account of the underlying biology of *in vitro* cell growth, and drug pharmacology.
3. Sensitivity, specificity, and predictive values are commonly used summary indices of *in vitro* - *in vivo* associations, but arise from dichotomization of laboratory and clinical data with loss of information. Objective criteria are needed for *in vitro* outcomes; and, *in vivo* outcomes should be chosen to reflect survival advantage.
4. Multivariate statistical techniques that assess whether *in vitro* assays provide additional information for predicting clinical outcome with other available prognostic criteria can be useful. However, these analyses presuppose adherence to statistical criteria for experimental design (including adequate sample sizes) and require prospective evaluation. Extrapolation of results from one clinical setting to another should be done with caution.

Clinical oncologists are presented with a wide variety of human tumors, and must select among many modalities for preferred treatment regimens. Their decisions may be enhanced by *in vitro* tests that can potentially predict *in vivo* chemotherapeutic response. The goals of *in vitro* testing are, in particular, to increase the likelihood of helping or curing the patient, and to avoid unnecessary toxicity by administration of ineffective agents; and, in general, to develop rapid and effective preclinical screens for new drugs. Shortcomings of available systems arise from technical problems in *in vitro* drug assessment, and incomplete understanding of biological factors determining drug response. The topic of *in vitro* drug assessment has been comprehensively and judiciously reviewed elsewhere, by Weisenthal[20], and it is not our intention to survey the field anew. Instead, our purpose is to emphasize that the application of rigorous statistical criteria to the design and analysis of *in vitro* assay techniques can lead to more reliable and dispassionate assessments of the utility of *in vitro* tests, of their interpretation and application in clinical oncology. For concreteness, we shall focus on the human tumor colony forming assay (Salmon et al[15]), though our principles apply more generally to other systems as well.

SOME STATISTICAL ISSUES

We begin with the fundamental scientific issue of reproducibility. As adjudged statistically, the assessment of any prognostic test entails an analysis of its variability. This is measured by its reproducibility and determined from independent test repetitions performed and interpreted under identical conditions. Clearly, low variability, or equivalently high reproducibility, is most desirable. In most published studies of *in vitro* drug testing, sufficient information is rarely given to pass judgment on this issue. For example, the distribution of colony numbers in a stem-cell assay might be expected to follow a Poisson distribution, due to random seeding of a large number (N) of cells having an inherently low prior probability (p) of colony formation. Using the Poisson assumption, one would expect $\lambda = Np$ colonies to be formed per plate with a statistical standard deviation of $\sqrt{\lambda}$. Furthermore, one should observe a linear relationship between cell numbers plated and colonies formed. If the Poisson assumption is valid, a colony assay technique is reproducible if the statistical variability implied by the assumption is sufficient to account for observed biological variability with independent repetitions. For example, a colony assay performed in triplicate and yielding colony numbers of 25, 30 and 35 might be considered eminently reproducible, as the observed standard deviation of 5 with a sample mean of 30 is within the biological variability predicted by a Poisson distribution. However, colony numbers of 20, 30, and 40, which maintain a sample mean of 30 but yield an observed standard deviation of 10, display variation significantly greater than that explained by a Poisson distribution, suggesting the assay is contributing extraneous variability to colony formation.

Reproducibility is a necessary but not sufficient property of a valid test, for reproducibility does not ensure test accuracy or precision. Tests that reproduce inaccurate or imprecise results are not valuable. Among variables to be considered, calculation of *in vitro* drug doses and exposures merits prominent attention. Many *in vitro* studies are conducted at a single drug concentration, such as 1/10 peak plasma levels, and using a one hr incubation (Salmon et al[15]). However, this procedure fails to consider pharmacokinetic data relating to:

a) the wide range of achievable levels of chemotherapeutic agents in humans (*in vivo* bioavailability)

b) quantitation of the proportion of target cells in various phases of the cell cycle, especially when measuring effects of relative phase specific drugs

c) parameters such as drug concentration x time, which may lead to different choice of doses *in vitro*

d) bioactivation requirements of certain drugs

Of particular concern is selection of a statistically precise endpoint measuring for *in vitro* assessment of cell death and correlation with *in vivo* effects. Two methods for *in vitro* assessment are commonly employed (Moon[10]):

a) an endpoint of percent reduction in colony formation (or, inhibition of isotope incorporation or dye exclusion), at a fixed drug dose

b) an endpoint determined by calculating the area under a survival vs. concentration curve

[Operationally, these endpoints may be equivalent: a correlation coefficient of 0.91 between them has been reported by Moon et al[11] for a series of 156 solid tumor patients.]

Aside from the obvious biological uncertainties, certain statistical issues remain. Ratio estimates (N/D), such as percent reduction, are notoriously unstable statistics: One must assume either that the variation in N and D is small, or if the variation in D is large, that the ratio N/D remains constant over a broad range of D's. Neither assumption is altogether justified with most assays. Instead, one might use the Poisson distribution of colony or cell growth to devise more reliable statistical estimates for differences in colony formation attributable to drug exposure. For example, suppose a cell colony assay yields a mean of 60 colonies per plate with triplicate controls, and a mean of 48 colonies in triplicate plates subsequent to drug exposure. Though the observed reduction in colony formation of 20% seems modest, a Poisson test for difference in rate of colony formation is statistically significant, at the $p = .05$ level (Armitage[1]). Therefore, to ascribe a negative *in vitro* result on the basis of the relatively low reduction in colony formation might be misleading, and cause a potentially useful drug to be dismissed from consideration.

[If for purposes of simplicity it is deemed desirable to dichotomize *in vitro* results, more objective statistical criteria are available. With the Poisson nature of colony formation, significant differences in colony growth between control and drug-exposed plates may be judged by the Poisson index of dispersion test, or by the likelihood ratio test, and used to assess *in vitro* response. Indeed, an early paper by Blackett[2] does exploit the Poisson nature of colony formation to examine the statistical accuracy of cell colony assays. His techniques can readily be extended to the analysis of *in vitro* outcomes.]

Note that percent reduction endpoints address the basic question of *in vitro* tumor sensitivity by assessing drug response only at a single point, e.g., a prespecified drug concentration, along a dose response curve. An alternative, more informative, approach for assessing *in vitro* drug sensitivity is to characterize the overall dose response curve (which necessitates multiple assays at various concentrations). Colony formation should decrease with increasing drug dosage. The log cell kill hypothesis (Skipper et al[16]) suggests that dose response curves should be simple negative exponentials, which can be characterized mathematically by the slope of the log colony versus concentration line. This parameter might provide a useful index of tumor sensitivity to the drug: The slope should be large for highly sensitive tumors, but near zero for relatively insensitive ones. [An alternative index might be the reciprocal of the slope, or some value proportional to it, such as the D_{10} value used by McCulloch et al[8]]. Statistical regression methods can be used to estimate the slope and assess its significance. However, in some situations the negative exponential characterization of dose-response is inadequate. For example, when shoulders or plateaus are present in the log colony versus concentration curves, slope indices derived solely from statistical considerations should be interpreted with caution. Also, choice of drug concentrations for study may dramatically affect the shape of drug survival curves.

If an *in vitro* test is reproducible, its prognostic utility can be assessed using the statistical criteria of sensitivity, specificity, and predictive value. Moon et al[11] describe a typical mechanism for calculating sensitivity and specificity. In brief, a "representative" group of individuals for whom *in vitro* response and *in vivo* drug assessments are available is found, and they are classified as in Table 1.

In the table, the sensitivity of the *in vitro* test equals a/a + c, the proportion of the *in vivo* responders labelled positive (e.g. complete remission, remission duration, survival) by the test; the specificity of the *in vitro* test equals d/b + d, the proportion of the *in vivo* nonresponders labelled negative by the test.

When applied to the evaluation of *in vitro* assays, a number of implications, limitations, and pitfalls attend this procedure. As these issues are salient to most *in vitro* tests, they will be discussed in some detail:

1. There must exist well-defined, objective criteria for the dichotomization of *in vivo* response: e.g., positive response represented by complete or partial tumor remission, and negative response by tumor progression or failure to achieve remission. One might argue, because remission is not synonymous with cure, only the therapeutic sensitivity of cells with renewal capacity will influence long-term tumor control. Hence for some assays an *in vivo* index reflective of stem cell kill and long-term remission (e.g., long-term disease-free survival) might be more appropriately related to *in vitro* outcome.

2. It is important to note whether individuals chosen for the cross-classification represent a random sample from a larger, well-defined population. Restricted samples will result in predictive parameters which cannot be used for broader populations. Also, calculated indices of sensitivity and specificity are statistical estimates of binomial probabilities and are subject to

Table 1. Cross-classification of *in Vitro* and *in Vivo* Responses

	in vivo +	*in vivo* −	
in vitro +	a	b	a + b = total number of *in vitro* responders
in vitro −	c	d	c + d = total number of *in vitro* nonresponders
	a + c = total number of *in vivo* responders	b + d = total number of *in vivo* nonresponders	

statistical errors. For example, Von Hoff et al[18] report the sensitivity of their *in vitro* tumor colony forming assay as 15/15+2=.88, but do not report the rather substantial standard deviation of .08 that should be attached to this figure. [Their specificity index, 100/100+6=.94, has a lower standard deviation of .02].

3. A fundamental criticism of standard mechanisms for effecting *in vitro-in vivo* association is lack of an objective, a priori *in vitro* criterion for dichotomization. Typically, the "cutoff" point for ascribing a positive versus negative *in vitro* response is selected retrospectively to optimize performance of some statistical criterion (e.g., maximal chisquared value), or to achieve a "harmonious" balance between sensitivity and specificity. Such criteria are best described as operational, and imputing statistical significance to subsequent statistical procedures is fallacious. Miller and Siegmund[9] discuss this fact in greater detail.

4. Sensitivity and specificity are inherent properties of any test, and one might expect they would remain invariant in different clinical settings. Such will not be the case, however, if:

a) the reproducibility of the test in different clinical settings has not been established (e.g., against multiple tumor subtypes)

b) the individuals upon whom assessments of *in vitro* and *in vivo* responses were made in one setting are not representative of those found elsewhere

c) operational criteria had been used for the dichotomization of *in vitro* results, without subsequent prospective evaluation and validation.

Sensitivity and specificity do not address the ultimate clinical question of whether an *in vitro* test can predict *in vivo* response. The predictive value of a positive test is the proportion of those with a positive test who respond *in vivo*, and the predictive value of a negative test the proportion of those with a negative test who do not respond. In a recent prospective assessment of a colony forming assay for solid tumors, Von Hoff et al[19] report a positive predictive value of 60% and a negative predictive value of 85% in a group with mostly *in vivo* nonresponders. As a laboratory test, these figures seem rather low to merit adoption of the assay.

Note that predictive value, unlike sensitivity and specificity, is dependent upon anterior probabilities, e.g., the prior probability of *in vivo* response in a given patient population. Suppose, for example, that a certain *in vitro* assay has a sensitivity of 65% and a specificity of 95%. The test is used in prospective evaluation of 1000 patients, for whom the a priori response rate to single-agent anti-cancer treatment is 20%. One expects 200 *in vivo* responses, of which 0.65 × 200 = 130 would be identified by *in vitro* testing. Similarly, one expects 800 non-responders to vivo therapy of which 0.95 × 800 = 760 would be identified *in vitro*. These outcomes can be summarized as in Table 2:

Table 2. Expected Outcome in Prospective Evaluation of 1,000 Patients with 20% a Priori Response Rate, If the *in Vitro* Assay has a Sensitivity of 65% and a Specificity of 95%

	in vivo +	*in vivo* −	Total
in vitro +	130	40	170
in vitro −	70	760	830
Total	200	800	1000

The predictive value of a positive test (the proportion of individuals with a positive *in vitro* test who are *in vivo* "responders") is therefore 130/(130+40)=0.76. The predictive value of negative test (the proportion of individuals with a negative test identified as non-responders *in vitro*) is 760/(760+70)=0.92. More generally, if p denotes the a priori probability of *in vivo* response in a particular population to be screened, with an *in vitro* assay having a sensitivity of S and specificity C, it follows from Bayes' theorem (Feller[4]) that the predictive value of a positive test PV+ (Vecchio[19]) is given by:

$$(1)\ PV+ \ = \ \frac{pS}{pS\ +(1-P)(1-C)}$$

and the predictive value of a negative test, PV−, is

$$(2)\ PV- \ = \ \frac{(1-p)C}{(1-p)C\ +\ p(1-S)}$$

The value of p depends on the particular patient population studied, but influences PV+ and PV−

dramatically. If testing a drug or tumor type with low overall clinical response rate, i.e., p near 0, PV− will tend to be near unity regardless of the test's sensitivity. The test would need a very high specificity to be clinically useful, that is, be able to predict *in vivo* response. Conversely, if testing a drug or tumor with a high over all response rate, i.e., p near unity, PV+ will also be near unity regardless of the test's specificity, and very high sensitivity would be needed for a clinical useful test (e.g., prediction of non-response *in vivo*) (Rozencweig and Staquet[14]). Extrapolating the predictive values of an assay from one clinical setting to another should also be done with extreme caution. Values of sensitivity and specificity (reflecting laboratory variation), and prior probabilities of *in vivo* response (reflecting the patient population) can be dissimilar in various clinical settings.

A more valid assessment of predictive value is whether the assay provides further information in addition to other available prognostic criteria such as age, sex, tumor type, histologic status, and prior therapy. Multivariate statistical procedures may be used to predict *in vivo* response from sets of prognostic variables; these procedures include discriminant analysis, logistic regression, and recursive partitioning. These three methodologies yield classification rules ascribing a particular *in vivo* response according to an individual's prognostic variables. Since the classification rules are essentially assignment procedures in a statistical decision-making process, it is useful and informative to assess accuracy of the decision rules. A widely used measure is the error rate, or misclassification rate, i.e., the probability of assigning an individual to the wrong *in vivo* outcome with a particular classification rule. These methodologies generally produce optimistic results when used retrospectively to classify the same cases from which the classification rules were computed. That is, apparent error rates generally underestimate true classification rates and present an overly optimistic picture; classification rules determined using a particular sample should perform better with that sample that with a new study group. A prospective evaluation of the classification rules provides a more accurate assessment and comparison of their error rates.

Alternatively, one might consider a continuous variable such as length of survival as the *in vivo* endpoint of interest, and thereafter assess the usefulness of the various prognostic criteria using the Cox regression model. McCulloch et al[8] and Curtis et al[3] incorporated both logistic regression and the Cox regression technique in a rigorous analysis of remission and survival in acute non-lymphoblastic leukemia, and found that *in vitro* drug sensitivities were not significantly related to either outcome after adjustment for other prognostic variables. Their results cast some doubt upon the predictive value of *in vitro* drug tests in ANLL and need to be verified in other clinical settings.

Of related interest is the report by Von Hoff et al[19] on the results of a prospective clinical trial of a human tumor clonogenic assay. One should interpret the results of this trial with some caution: Von Hoff et al noted that the assay was workable for less than half of the patient population to which it was applied; hence the possibility of selection bias with the assay arises. Hug et al[5] examined the issue of selection bias in a different series of patients; and, Johns and Mills[6] noted that cloning efficiency might be of prognostic significance.

A chi-squared statistic (X^2)(or, Fisher's exact test) based on the *in vitro* - *in vivo* cross-classification table is sometimes reported as a measure of "correlation" between *in vitro* and *in vivo* outcomes (Park et al[12]). However, this is less useful than examination of the individual indices of sensitivity, specificity, and predictive value. The summary X^2 statistic essentially provides a test of the null hypothesis that sensitivity + specificity = 1. A significant X^2 statistic reflects a value of sensitivity + specificity that is statistically greater than or less than 1, but provides no further information, thereby blurring the distinction between these two indices. Similarly, Park et al[12] report a measure of overall "accuracy" or "correlation" between *in vitro* and *in vivo* outcomes as the number of "correct" *in vitro* predictions of *in vivo* outcome divided by the total number of predictions. In the above notation, this index of validity is equal to pS + (1-p)C, and can assume any value between S and C as p varies between 0 and 1. As with X^2, the index of validity is less informative than examination of the individual indices of sensitivity and specificity.

More important, however, the use of summary indices is predicated on dichotomization of *in vitro* and *in vivo* outcomes, which is an extreme oversimplification of laboratory and clinical results.

One infers from the forgoing discussion that complex multivariate statistical techniques are required to relate *in vivo* outcome or survival to available explanatory variables. When implementing procedures such as discriminant analysis, logistic regression, recursive partitioning, or Cox regression, there is no necessity to dichotomize the *in vitro* assay result; indeed, such dichotomization always implies an unfortunate loss of information. Nor is there any necessity to restrict consideration to a single value from the *in vitro* assay: Various indices such as line slopes and levels of concentration-dependent killing may be assessed simultaneously. Of special interest is the shape of the *in vitro* dose response curve (and hence the inadequacy of assessing *in vitro* response at one particular point along this curve): An exponential curve might indicate that cellular sensitivity is relatively homogeneous within the tumor, whereas the presence of resistant cell sublines (heterogeneity) among tumor cells might be revealed by plateaus at high drug levels, or a shoulder at low drug doses. Such information has clear impact on clinical outcome, and can readily be incorporated into the multivariate methodologies previously described.

Statistical precision of both univariate statistics, such as sensitivity, specificity, and predictive value, and multivariate procedures such as the classification rules derived from discriminant analysis, logistic regression, or recursive partitioning, improve with increasing sample sizes. Indeed, with small sample sizes, standard errors of estimates can be so large that estimates themselves become misleading. Multivariate techniques are especially prone to possibly aberrant results with small sample sizes. Thus, for example, a recent study by Lihou and Smith[7)] was based on only 19 patients, and precision of multivariate classification rules derived therefrom is of concern.

ACKNOWLEDGEMENTS

This research was supported from grants R01-CA26666 and K04-CA00687 from the U.S. National Cancer Institute.

REFERENCES

1. **Armitage, P.** 1971. Statistical Methods in Medical Research p. 138-140. John Wiley, New York.
2. **Blackett, N.M.**1974. Statistical accuracy to be expected from cell colony assays; with special reference to the spleen colony assay. Cell Tissue Kinet. **7**: 407-412.
3. **Curtis, J.E., Messner, H.A., Hasselback, R., Elhakim, T.M. and McCulloch, E.A.** 1984. Contributions of host- and disease- related attributes to the outcome of patients with acute myelogenous leukemia. J. Clin. Oncol. **2**: 253-264.
4. **Feller, W.** 1957. An Introduction to Probability Theory and Its Applications, p. 114. Vol. 1, Second Edition. John Wiley, New York
5. **Hug, V., Thames, H., Johnston, D., Blumenschein, G., Drewinko, B. and Spitzer, G.** 1984. The true predictive value of the human tumor stem cell assay: does a workable assay select for treatment responders? J. Clin. Oncol. **2**: 42-45.
6. **Johns, M.E. and Mills, S.E.** 1983. Cloning efficiency. Cancer **52**: 1401-1404.
7. **Lihou, M.G. and Smith, P.J.** 1983. Quantitation of chemosensitivity in acute myelocytic leukemia. Br. J. Cancer **48**: 559-567.
8. **McCulloch, E.A., Curtis, J.E., Messner, H.A., Senn, J.S. and Germanson, T.P.** 1982. The contribution of blast cell properties to outcome variation in acute myeloblastic leukemia. Blood **59**: 601-608.
9. **Miller, R. and Siegmund, D.** 1982. Maximally selected chi square statistics. Biometrics **38**: 1011-1016.
10. **Moon, T.E.** 1980. Quantitative and statistical analysis of the association between in vitro and in vivo studies. p. 209-221. In Cloning of Human Tumor Stem Cells, S.E. Salmon, Ed., Alan R. Liss, New York.
11. **Moon, T.E., Salmon, S.E., White, C.S., Chen, H.-S. G., Meyskens, F.L., Durie, B.G.M. and Alberts, D.S.** 1981. Quantitative association between the in vitro human tumor stem cell assay and clinical response to cancer chemotherapy. Cancer Chemother. Pharmacol. **6**: 211-218.
12. **Park, C.H., Amare, M., Morrison, F.S., Maloney, T.R. and Goodwin, J.W.** 1982. Chemotherapy sensitivity assessment of leukemic colony-forming cells with in vitro simultaneous exposure to multiple drugs: clinical correlations in acute nonlymphocytic leukemia. Cancer Treat. Rep. **66**: 1257-1261.
13. **Park. C.H., Wiernik, P.H., Morrison, F.S., Amare, M., Van Sloten, K. and Maloney, T.R.** 1983. Clinical correlations of leukemic clonogenic cell chemosensitivity assessed by in vitro continuous exposure to drugs. Cancer Research **43**: 2346-2349.
14. **Rozencweig, M. and Staquet, M.** 1984. Predic-

tive tests for the response to cancer chemotherapy: limitations related to the prediction of rare events. Cancer Treat. Rep. **68**: 611-613.

15. **Salmon, S.E., Hamburger, A.W., Soehnlen, B., Durie, B.G., Alberts, D.S. and Moon, T.E.** 1978. Quantitation of differential sensitivity of human tumor stem cells to anticancer drugs. N. Engl. J. Med. **298**: 1321-1327.
16. **Skipper, H.E., Schabel, F.M. and Wilcox, W.S.** 1964. Experimental evaluation of potential anticancer agents. XII. On the criteria and kinetics associated with "curability" of experimental leukemia. Cancer Chemother. Rep. **35**: 1-111.
17. **Vecchio, T.J.** 1966. Predictive value of a single diagnostic test in unselected populations. N. Engl. J. Med. **274**: 1171-1173.
18. **Von Hoff, D.D., Casper, J., Bradley, E., Sandbach, J., Jones, D. and Makuch, R.** 1981. Association between human tumor colony-forming assay results and response of an individual patient's tumor to chemotherapy. Amer. J. Medicine **70**: 1027-1032.
19. **Von Hoff, D.D., Clark, G.K., Stogdill, B.J., Sarosdy, M.F., O'brien, M.T., Casper, J.T., Mattox, D.E., Page, C.P., Cruz, A.B. and Sandbach, J.F.** 1983. Prospective clinical trial of a human tumor cloning system. Cancer Research **43**: 1926-1931.
20. **Weisenthal, L.M.** 1981. In vitro assays in preclinical antineoplastic drug screening. Semin. Oncol. **8**: 362-376.