Alternative Tests: Carcinogenesis as an Example

Bernard Schwetz and David Gaylor

U.S. Food and Drug Administration/National Center for Toxicological Research, Jefferson, Arkansas

Acceptance of new tests that are alternatives to currently used toxicology tests is a topic of considerable importance in the field of toxicology. Carcinogenicity testing today normally includes 2-year studies in rats and mice of both sexes, following widely accepted procedures for husbandry; selection of dose levels; pathology and toxicity observations; and statistical interpretation of tumor data. These studies are usually preceded by tests for genetic toxicity and subchronic toxicity studies to select dose levels for the 2-year studies. Although these data are used for quantitative risk assessment, the mechanistic basis for effects is usually unknown. The series of studies is very expensive and requires 5 years or more to conduct. Alternative approaches are being developed that would provide more mechanistic information and hopefully would permit decisions to be made about carcinogenic potential without the need to conduct 2-year studies in rats and mice of both sexes. Decisions could be based on a profile of data rather than on the result of one test. Procedures for regulatory acceptance of new approaches for carcinogenicity testing are critical to future progress. — *Environ Health Perspect* 106(Suppl 2):467–471 (1998). *http://ehpnet1.niehs.nih.gov/docs/1998/Suppl-2/467-471schwetz/abstract.html*

Key words: carcinogenesis, alternative tests, regulations

Introduction

Acceptance of alternative tests represents an opportunity to introduce new test methods into the toxicologist's armamentarium that will eventually reduce our dependence on resource-intensive whole animal tests. New test designs must be sensitive to the desire for mechanistically based data and the "three Rs"—reduction of the number of animals, refinements to enhance the well-being of animals, and replacements that do not use whole animals or lower species. This paper presents a strategy for using alternative tests for carcinogenesis screening. The term carcinogen is used here in the broad sense as an agent capable of increasing the incidence of malignant neoplasia after exposure. The concepts underlying the strategy apply to other areas of toxicology in which commonly used *in vivo* tests can be potentially replaced by alternative tests.

It has been suggested for some time that the use of the 2-year rodent bioassay as a screen, not as a definitive study for carcinogenic potential, but only as a screen. The reason for this has been that the 2-year bioassay as currently conducted is approximately one million dollars per study (single sex–species combination) and frequently gives equivocal results. Consequently, this is an expensive and time-consuming screen that often does not give definitive answers. This recommendation to seek alternatives was based on the assumption that the 2-year bioassay would be replaced with better test systems that had the desirable characteristics of being cheaper and faster, using fewer animals, and providing the appropriate sensitivity and specificity desired of a screen for carcinogenic potential. For many years, the desire to replace the 2-year rodent bioassay with other tests was a matter of talk with very little action to accomplish that objective. During the past 5 years, however, there has been clear movement toward acceptance of some alternatives to the 2-year study. Presently, one of the problems is the segregation of toxicologists into three groups in response to a movement toward alternative test systems. First, are those who now recommend no change and a continuation of 2-year studies in two species of both sexes when there is a need for carcinogenicity data. A second group supports the proposal of the International Committee for Harmonization for drugs (*1*); that is, when carcinogenicity data are required, the studies would be conducted in one rodent species and be supplemented by other data. The default rodent species would be the rat. A third group recommends simply conducting carcinogenicity studies in one species of rodent, preferably the rat, with no requirement for other data.

None of these choices is clearly right or wrong. One of the problems with our approach over the last two decades is that we have accepted the standard 2-year bioassay, typified by the study design that has been used by the National Toxicology Program (NTP) in the past, as the standard study design for all chemicals, reflecting the philosophy that "one size fits all." As we learn more about the modes of action of carcinogens, it becomes clear that some other alternative must be considered and that the selection of studies must be tailored to the individual chemical. The choice of tests should be based on the chemical properties of the substances as well as on the available biologic and toxicologic data. The nature and amount of human exposure involved must also be considered. For example, different tests might be warranted for a food additive, pesticide, or a substance for which there is occupational exposure compared to a drug with a very limited population and duration of exposure. The test profile might be different for a low-level contaminant in a product compared to the active ingredient. Rather than simply categorize chemicals as carcinogens or noncarcinogens, Ames et al. (*2*) suggest examining the ratio of anticipated human exposure to a dose producing tumors in rodents.

In this paper we present a strategy to reduce dependence on 2-year studies for assessing carcinogenic potential and a review of the 2-year rodent study accepted as the "gold standard." Progress in accepting alternatives to the 2-year studies depends, first, on acceptance of a new test paradigm and, second, on a need to move away from past tradition of requiring 2-year studies in both sexes of two species every time carcinogenicity data are required. These two steps are not as interdependent as the idea of replacement of a gold standard might suggest.

This paper was prepared as background for the 13th Meeting of the Scientific Group on Methodologies for the Safety Evaluation of Chemicals (SGOMSEC): Alternative Testing Methodologies held 26–31 January 1997 in Ispra, Italy. Manuscript received at *EHP* 9 May 1997; accepted 11 September 1997.

Address correspondence to Dr. B.A. Schwetz, FDA/National Center for Toxicological Research, 3900 NCTR Road, Jefferson, AR 72079-9502. Telephone: (870) 543-7517. Fax: (870) 543-7576. E-mail: bschwetz@nctr.fda.gov

Abbreviations used: U.S. FDA, U.S. Food and Drug Administration; NTP, National Toxicology Program.

Review of the 2-Year Rodent Studies

Because acceptance of alternatives to 2-year studies will require change, it is important to identify the constraints of such a change. The precedent of having required two-species, 2-year rodent studies has provided us with a certain level of comfort in knowing what to require when carcinogenicity data are desired. In addition to the comfort that rests with that tradition, considerable importance has been attributed to the historical database. Also, it is human nature that significant changes are hard to effect. Both the regulated industry and regulators know what to expect under the current paradigm, what to expect in terms of testing when it is required and what the interpretation will be depending on the outcome of the study. As a result of these constraints, change has been slow.

Change is required, however. There are a number of often-cited reasons for change that are obvious, such as the cost of 2-year studies, the use of large numbers of animals, the duration required to conduct the studies, and the uncertainty of the outcome in terms of equivocal answers or extrapolation of results to humans. For example, because of the high background incidence of liver tumors in male $B6C3F_1$ mice, discussions continue on the relevance of this tumor for humans.

There are some less obvious reasons for change that must also be pointed out. Two-year rodent studies are empirical in nature; the assumption is that the development of most tumors, even by unknown mechanisms, is predictive of some tumorigenic potential in humans. Some notable exceptions are kidney tumors formed in male rats due to an accumulation of the $\alpha_2\mu$-globulin protein and bladder tumors due to the formation of crystals in the urine.

The historical database we have relied on is not as useful as was assumed. For example, there has been genetic drift in many of the strains of animals believed to be genetically pure strains. The observation that the body weight of certain test animals has increased progressively during the past few years has complicated the use of control data (*3*). Animals live a shorter time and the profile of tumors is different. Therefore, control data generated 15 years ago, when mean body weights were considerably less, cannot be the basis for comparisons to studies conducted today in heavier animals (*4*). Consequently, we have moved away from using historical databases to the use of contemporary databases, eliminating control data from earlier years. The primary emphasis is now given to concurrent control data, again recognizing the fact that control data from previous years are less valuable than concurrent control data.

We have been concerned about the predictivity of 2-year bioassays over the last decade as our experience and knowledge have expanded. This issue must be addressed as we consider the benefit of alternative test systems. New test systems reflect new scientific developments, that is, new models that are mechanistically based and reflect the increasing diversity of mechanisms that contribute to the development of the carcinogenic response. Such information is not obtained from 2-year rodent studies as they have been conducted in the past.

One of the evaluative measures of the 2-year study as a gold standard is the predictiveness observed between rats and mice. The species concordance has been evaluated and reviewed by many investigators, including DiCarlo (*5*), who reviewed the results of 138 NTP bioassays and found that the rat–mouse concordance was 75%. More recently, Huff et al. (*6*) reviewed the concordance of 379 NTP bioassays and found a 74% level of rat–mouse concordance. Contrera et al. (*7*) reviewed the results of 282 drugs in the U.S. Food and Drug Adminstration (U.S. FDA) database and again found a 74% rat–mouse concordance. The importance of this observation is that the U.S. FDA database consists primarily of studies in Sprague-Dawley rats and Swiss Webster-derived mice; the NTP database is based primarily on results from Fischer 344 rats and $B6C3F_1$ mice. The similarity and concordance in these databases lends validity to the observation that the numbers reflect the response of rodent species. It is difficult to imagine a better predictor of rat carcinogenicity than mouse carcinogenicity or vice versa. The overall concordance, however, between rats and mice has been identified repeatedly as about 75%.

Potency is another measure of the usefulness of the 2-year rodent studies. Carcinogenic potency is defined as the lifetime risk per unit of average daily exposure. Exposure is generally expressed as milligram per kilogram body weight or concentration in food, water, or air. Target tissue concentrations are seldom available. Crouch and Wilson (*8*) and Crouch (*9*) concluded that there was good species correlation for those chemicals that were carcinogenic in both rats and mice. The potencies of the two species were generally within a factor of 20 of each other. Note that this is true only for chemicals that were carcinogenic in both species.

Using the carcinogenic potency database compiled by Gold and co-workers (*10*) on 770 compounds, Gaylor and Chen (*11*) showed good agreement for carcinogenic potency, on the average, among rats, mice, and hamsters for various routes of exposure. However, variability was substantial, with differences generally within a factor of 100. For 69 NTP chemicals that produced tumors in the same sex and tissue site, Chen and Gaylor (*12*) showed that carcinogenic potencies for rats and mice were generally within a factor of 40. Despite the variability, it has been argued that the correlation of potencies between rats and mice supports extrapolation to humans. However, the studies were only conducted on concordant chemicals, i.e., chemicals that produce tumors in both rats and mice.

Tennant et al. (*13*) prospectively predicted the carcinogenic status of 44 NTP chemicals based upon chemical structure, *Salmonella* assay results, dose level, and subchronic pathology. Wachsman et al. (*14*) presented the results of the Tennant et al. (*13*) predictions, along with the predictions of seven other groups, with the outcome for 40 of the chemicals. There was substantial agreement among the eight predictive systems on 14 of 40 (35%) chemicals that were clearly positive or negative. Human expert systems performed better than computer-based systems. There was a good correlation between electrophilicity and carcinogenicity. A high percentage of equivocal results in the 2-year bioassays makes it difficult to validate predictive systems. Further, Ashby and Tennant (*15*) concluded from this exercise that the integration of different predictive techniques is preferable to the use of single techniques. They concluded that carcinogenic predictivity appears to be limited to less than 80%. Part of the problem is that the 2-year bioassay cannot detect weak carcinogens, and high doses may produce carcinogenicity by indirect mechanisms that would not be predicted at lower dose levels. That is, the 2-year bioassay cannot always accurately determine the carcinogenicity of a chemical. Nonetheless, many accept the 2-year bioassay as the best available test for carcinogenicity at the present time, despite the fact that the 2-year bioassay has a relatively high rate of negative conclusions (*16*). There are questions about the relevancy of certain tumor types observed in rodents to humans. This is particularly

true for very high dose levels of exposure. Ashby and Purchase (*17*) suggest that with adequate testing in long-term bioassays, few chemicals are likely to be considered noncarcinogenic. Haseman (*18*) discusses the low sensitivity of standard bioassays in detecting weak carcinogens. Based on an analysis of NTP studies by Gaylor (*19*), if 100 animals had been used per dose group rather than 50 animals, with the resulting increased power of detection and the same incidence rates, it appears that at least 70% of NTP chemicals would be considered animal carcinogens. These results suggest that the standard 2-year rodent bioassay employing high-dose levels may, in effect, be a long-term toxicity test in which cancer is often one of the biologic manifestations, limiting its usefulness as a good screen for carcinogenicity.

Our scientific knowledge about the mechanisms of carcinogenesis is far different today than it was in the 1970s when the 2-year bioassay was first adopted as a routine screen. In addition to information from genetic toxicity assays and structural alert information, it is apparent today that a broad range of toxicologic and genetic toxicity data would be helpful in supporting predictions about the probable carcinogenicity of a substance. Such information would include cell proliferation data, information on apoptosis, peroxisome proliferation capabilities, impact on hormonal profile, production of $\alpha_2\mu$-globulin, profile of metabolism, and other information.

In summary, with the 2-year bioassay, the rat–mouse prediction for carcinogenicity is about 75% accurate. The bioassay is poor at detecting the effect of weak carcinogensand is not useful for evaluation of the mechanisms of carcinogenicity, an array of information much more diverse than anticipated when the bioassay was first adopted over two decades ago. These limitations raise obvious questions about the usefulness of the bioassay as the gold standard from which to draw conclusions regarding alternatives to the 2-year bioassay, and must be considered as we move toward accepting alternatives to the standard rodent bioassay.

Alternative Strategies

Within the U.S. FDA, we have had considerable discussion over the past 3 years regarding the desire to develop a strategy for carcinogenicity testing that would permit regulatory decisions to be made on data sets that may not include the results of traditional 2-year studies in both sexes of two rodent species. Conclusions about carcinogenic potential should be based on a profile of toxicologic data, not just on a bioassay result. The weight of evidence is more important than an approach that depends on a decision tree. Conclusions should consider data that describe mechanisms of carcinogenesis. One such strategy might include use of data from new test systems such as the transgenic models TG:p53+/–, the p53 hemizygote, and the transgenic model TG:AC. The relevancy of the transgenic models to humans must be evaluated. Another test system currently being reevaluated is the newborn mouse assay.

A strategy that includes the use of these test systems is diagrammed in Figure 1. Assuming there is human exposure and therefore a desire to collect information about carcinogenic potential, one would base preliminary estimates on physical chemical properties of the substance, structural alert information, information from computer-based predictive systems, and the results of the genetic toxicity screen. On the basis of this information, one would conclude that a substance is either nongenotoxic or genotoxic. If the substance is strongly genotoxic, one could conclude that the chemical would be a carcinogen if tested in a 2-year study, or the alternate would be to proceed with the 2-year rodent study. However, because about one-third of mutagens are not carcinogenic in the standard bioassay and about one-third of carcinogens are not mutagenic in common tests, it probably would be prudent to

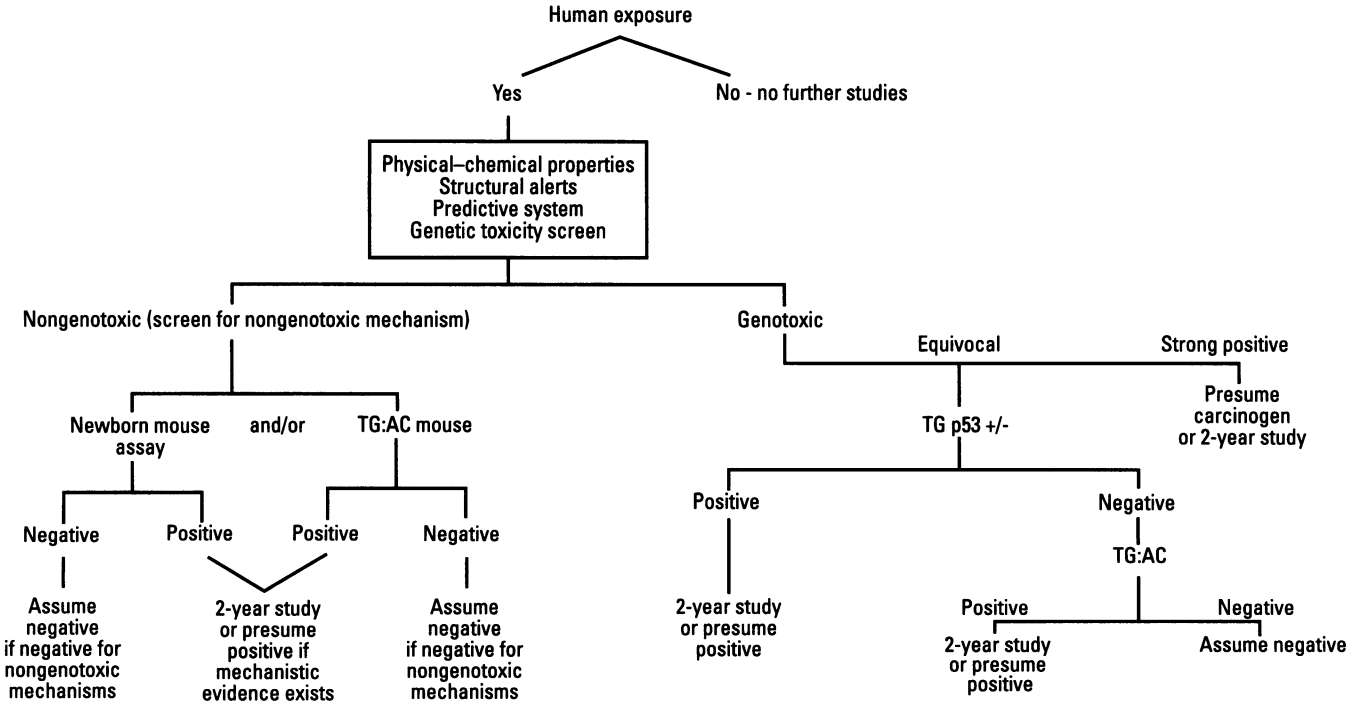

Figure 1. Example of a potential strategy.

proceed with further testing. For genotoxic chemicals, one might use the p53 hemizygote as the next level of screen. If this test is positive, one has the option of conducting a 2-year study to establish whether the chemical is carcinogenic, or of presuming that this genotoxicant is carcinogenic because it was positive in the p53 hemizygote, which is sensitive to genotoxic carcinogens. If the substance is negative in the p53 hemizygote, one might test in the TG:AC to evaluate mechanisms of carcinogenicity other than genotoxicity. If positive in this screen, one might assume that the substance is carcinogenic or conduct a 2-year study to clarify that potential. If the substance is negative in the p53 hemizygote and negative in the TG:AC strain, one might be able to assume that this substance is negative for carcinogenic potential.

For substances that are nongenotoxic, one might proceed to either the newborn mouse assay or to the TG:AC mouse. The newborn mouse assay is included here because of uncertainty at this point about its range of predictiveness. Prior studies have shown that this test is clearly sensitive to genotoxicants; studies to evaluate its ability to identify nongenotoxicants are currently in progress. Nonetheless, if a nongenotoxic chemical is positive in either the TG:AC or the newborn mouse assay, one has the option of conducting a 2-year study or presuming that the substance is positive, particularly if there is mechanistic evidence that the chemical might be a carcinogen by some nongenotoxic mechanism. If the nongenotoxic chemical is negative in the newborn mouse assay and/or the TG:AC mouse, one might assume that this chemical is not a carcinogen, particularly if an evaluation of mechanisms for nongenotoxic carcinogenicity reveals no evidence that this substance would be carcinogenic through nongenotoxic mechanisms. Because this test system provides tumor incidence rates, it is also possible to consider the ratio of doses in animals producing tumors to expected human exposures. This ratio provides a method for setting priorities for regulatory action and further testing.

This test strategy, which includes short-term *in vivo* screens for carcinogenicity, also depends heavily on accumulation of data that relate to nongenotoxic mechanisms of carcinogenicity. This includes information on mechanisms identified in Table 1 that include hormone modulation, perturbation of growth factors, changes in cell proliferation, apoptosis, and other mechanisms of carcinogenicity that do not depend upon primary damage to DNA.

Further comments are warranted about the feasibility of this test strategy for carcinogenicity. First, this is simply a conceptual framework from which a test strategy could evolve. The strategy is not ready for widespread implementation as a complete test strategy at the present time because some components have not been fully evaluated, and certainly not validated in the traditional sense of formal acceptance of a new test system. This raises the question, however, of what validation means when a test system is mechanism based. To expend considerable resources in multilaboratory studies conducted on large numbers of chemicals through blinded procedures to confirm that a mechanism-based test system does not detect carcinogens by other mechanisms does not seem to be a useful exercise. Some components of the strategy presented in this paper will likely be replaced or supplemented by others in the future. Thus, mention of the newborn mouse assay or specific transgenic systems is not an endorsement that these are test systems most likely to be in use several years into the future.

There must be acceptance of alternative test systems by scientists in regulatory agencies as well as in the regulated industry and in academia for these test systems to be important components of test systems in the future. The development of the Interagency Coordinating Committee on the Validation of Alternative Methods takes us a long way toward that requirement.

One way to facilitate the acceptance of new test systems would be to conduct studies in parallel, using the traditional 2-year bioassay together with new test systems. This has been proposed many times in the past and has not been implemented to an appreciable extent because of the significant costs involved. The NTP, however, is currently supporting the further evaluation of transgenic models including the p53 hemizygote, the TG:AC model, the *ras*H2 model, and the newborn mouse assay. Other models are being evaluated, including the Eμ-*pim*-1 model, a TGFα model, and a *Xeroderma pigmentosum* model. Still another multilaboratory evaluation is being conducted on several of these systems through efforts organized by International Life Sciences Institute (*20*).

Table 1. Measures of altered cell function.

Hormone modulation
Steroid (estrogen, androgen, retinoid)
Growth factor perturbation
Cell proliferation (mitogenic, cytotoxic)
Specific tissue responses
Bladder, stones
Liver, necrosis
Kidney, $\alpha_2\mu$
Forestomach
Inhibition of apoptosis
Specific mechanisms
β-Agonist, uterine tissue
H_2-antagonist, glandular stomach
Peroxisome proliferation
Cell-to-cell communication
P450 induction
Spindle fiber effects
Altered methylation status

One other factor that impacts on the acceptance of alternative test systems is the level of reliance that we have placed on the results of negative 2-year rodent studies in the past. Through use of these bioassays for over 25 years, we have derived a certain level of confidence in the significance of negative results in these tests. A major concern about acceptance of alternative test systems is how regulatory agencies will use either positive or negative results from these test systems. One must consider whether we rely too heavily on the results of 2-year studies compared to our fear of using the results of negative alternative studies.

Summary

The 2-year rodent bioassay as we have used it for the past 25 years has been very useful, but we recognize that it is not perfect. We suggest that the 2-year bioassay has limited usefulness as a standard for evaluation of alternative test strategies. New test strategies should be evaluated on their own merit, as the 2-year bioassay has been over the past 25 years. We are in a period of transition during which results of short-term tests, in conjunction with the results of 2-year bioassays, will permit us the opportunity to evaluate the performance of these new short-term tests. The recommendations of the International Committee on Harmonization to accept carcinogenicity data in one species and other test systems as an alternate to data from two rodent species are evidence that, at least for drugs, there is a desire to make changes in the requirement for carcinogenicity data.

REFERENCES AND NOTES

1. U.S. Food and Drug Administration. International Conference on Harmonization: draft guideline on testing for carcinogenicity of pharmaceuticals. Fed Reg 61(163):43298–43300 (1996).
2. Ames BN, Magaw R, Gold LS. Ranking possible carcinogenic hazards. Science 236:271–280 (1987).
3. Haseman JK, Rao GN. Effects of corn oil, time-related changes, and interlaboratory variability in tumor occurrence in control Fischer 344 (F344/N) rats. Toxicol Pathol 20:52–60 (1992).
4. Seilkop SK. The effect of body weight on tumor incidence and carcinogenicity testing in B6C3F1 mice and F344 rats. Fundam Appl Toxicol 24:247–259 (1995).
5. DiCarlo FJ. Carcinogenesis bioassay data: correlation by species and sex. Drug Metab Rev 15:409–413 (1984).
6. Huff J, Cirvello J, Haseman JK, Bucher J. Chemicals associated with site-specific neoplasia in 1394 long-term carcinogenesis experiments in laboratory rodents. Environ Health Perspect 93:247–270 (1991).
7. Contrera JF, Jacobs AC, DeGeorge JJ. Carcinogenicity testing and the evaluation of regulatory requirements for pharmaceuticals. Regul Toxicol Pharmacol 25:130–145 (1997).
8. Crouch E, Wilson R. Interspecies comparison of carcinogenic potency. J Toxicol Environ Health 5:1095–1118 (1979).
9. Crouch EAC. Uncertainties in interspecies extrapolations of carcinogenicity. Environ Health Perspect 50:321–328 (1983).
10. Gold LS, de Veciana M, Backman GM, Magaw R, Lopipero P, Smith M, Blumenthal M, Levinson R, Gerson J, Bernstein L et al. Chronological supplement to the Carcinogenic Potency Database: standardized results of animal bioassays published through December 1982. Environ Health Perspect 67:161–200 (1986).
11. Gaylor DW, Chen JJ. Relative potency of chemical carcinogens in rodents. Risk Anal 6:283–290 (1986).
12. Chen JJ, Gaylor DW. Carcinogenic risk assessment: comparison of estimated safe doses for rats and mice. Environ Health Perspect 72:305–309 (1987).
13. Tennant RW, Spalding J, Staisiewicz S, Ashby J. Prediction of the outcome of rodent carcinogenicity bioassays currently being conducted on 44 chemicals by the National Toxicology Program. Mutagenesis 5:3–14 (1990).
14. Wachsman JT, Bristol DW, Spalding J, Shelby J, Tennant, RW. Predicting chemical carcinogenesis in rodents. Environ Health Perspect 101:444–445 (1993).
15. Ashby J, Tennant RW. Prediction of rodent carcinogenicity for 44 chemicals: results. Mutagenesis 9:7–15 (1994).
16. Haseman JK. Statistical issues in the design, analysis and interpretation of animal carcinogenicity studies. Environ Health Perspect 58:385–392 (1984).
17. Ashby J, Purchase IFH. Will all chemicals be carcinogenic to rodents when adequately evaluated? Carcinogenesis 8:489–495 (1993).
18. Haseman JK. A reexamination of false-positive rates for carcinogenesis studies. Fundam Appl Toxicol 3:334–339 (1983).
19. Gaylor DWG. Unpublished data.
20. International Life Sciences Institute. ILSI HESI mobilizes rapid response to need for data on alternative approaches in carcinogenicity testing. ILSI News 14(5):3 (1996).