Risk Analysis, Vol. 10, No. 4, 1990

Comment

The Importance of Being Quantitative When Crying "Fallacy"[1]

Gay Goodman[2,3]

Wartenberg and Gallo[1] find fault with Ames *et al.*[2] for using the rank order of rodent TD_{50}s to predict the rank order of carcinogen hazards to humans when exposures are at much lower levels than the doses typically used in the rodent bioassays. This is certainly a topic that merits serious investigation, but Wartenberg and Gallo do not offer any new perspectives or much analysis of existing information. Rather than provide a thorough investigation of the problem, their sole purpose seems to have been to refute a paper that they apparently did not read carefully enough.

The authors call Ames *et al.*[2] to task for proposing "a new model" in which the rodent TD_{50} is used as the basis for evaluating human carcinogenic hazards. But the TD_{50} or a similar high-dose measure of potency is already widely used for risk assessment, and Ames *et al.*[2] are quick to point out that this is an inadequate, if sometimes expedient, approach:

> Extrapolation from the results of rodent cancer tests done at high doses to effects on humans exposed to low doses is routinely attempted by regulatory agencies when formulating policies attempting to prevent future cancer. There is little sound scientific basis for this type of extrapolation, in part due to our lack of knowledge about mechanisms of cancer induction, and it is viewed with great unease by many epidemiologists and toxicologists [Ref.]. Nevertheless, to be prudent in regulatory policy, and in the absence of good human data (almost always the case), some reliance on animal cancer tests is unavoidable. (p. 271)

Ames *et al.* are not arguing for more widespread use of rodent potencies measured at high doses (e.g., the TD_{50}) in determining human risk. However, they suggest that since high-dose measurements are the only carcinogenicity data available for most chemicals, then the TD_{50} may be used, in combination with probable human exposure levels, to establish a rough estimate of the rank order of cancer hazards. What is original in the paper by Ames *et al.* is *not* the use of the rodent data in itself; rather, it is the introduction of a simple method to combine human exposure information with rodent carcinogenicity data for obtaining an index of expected human risk (i.e., the HERP).

The essential argument of Wartenberg and Gallo is that the high-dose potency or even the shape of the dose–response curve at high doses does not provide information about the dose–response at low doses, and in particular does not tell you whether the curve deviates sharply from linearity at low doses. Many competent researchers (e.g., Swenberg *et al.*[3]) are aware of this difficulty and have addressed the issue recently. Ames *et al.*[2] state:

> It would be a mistake to use our HERP index as a direct estimate of human hazard. First, at low dose rates human susceptibility may differ systematically from rodent susceptibility. Second, the general shape of the dose–response relationship is not known. A linear dose–response has been the dominant assumption in regulating carcinogens for many years, but this may not be correct. If the dose–responses are not linear but are actually quadratic or hockey-stick-shaped or show a threshold, then the actual hazard at low dose rates might be much less than the HERP values would suggest. An additional difficulty is that it may be necessary to deal with carcinogens that differ in their mechanisms of action and thus in their dose–response relationship. We have therefore put an asterisk next to HERP values for carcinogens that do not appear to be active through a genotoxic (DNA damaging or mutagenic) mechanism [Ref.] so that comparisons can be made within the genotoxic or nongenotoxic classes. (p. 272)

Since information pertaining to the really interesting, low-dose responses is statistically unattainable in experiments with 100 or fewer animals per dose group, the high-dose potency is often the only parameter available to risk assessors. Most workers in this area would probably agree that bioassay design could be modified so as to lessen the high-dose/low-dose extrapolation problem—for instance by performing time-to-tumor measurements and by looking for preneoplastic changes in animals treated at low doses. However, Wartenberg and

[1] Received Janauary 8, 1990.

[2] Harvard University, Department of Physics, and Energy and Environmental Policy Center, Cambridge, Massachusetts 02138.

[3] Current address: Gradient Corporation, 44 Brattle Street, Cambridge, Massachusetts 02138.

0272-4332/90/1200-0619$06.00/1

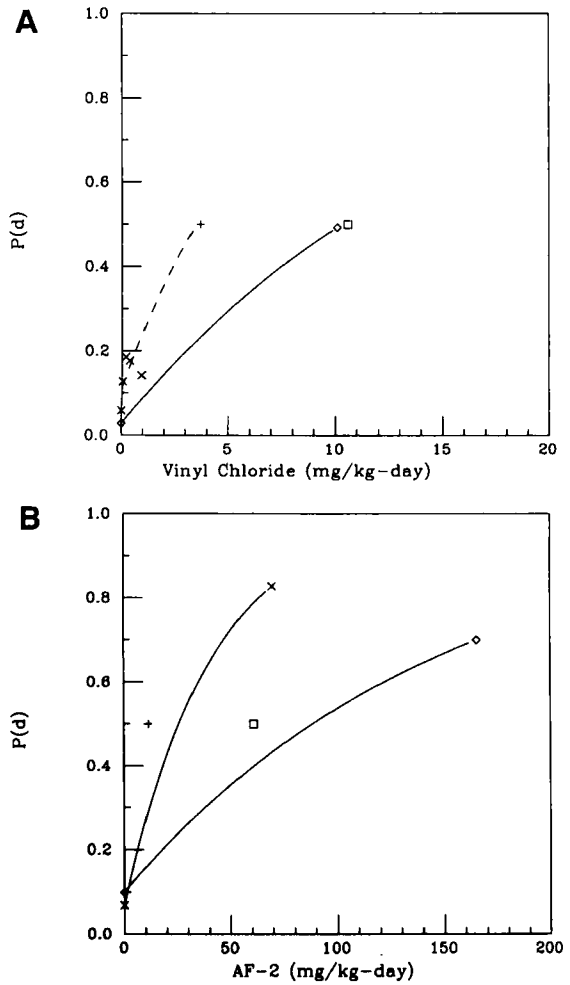

Fig. 1. Dose–response data and TD_{50}s at the most sensitive site and best-fit curves for vinyl chloride (A) and AF-2 (B). Data and TD_{50}s are from Gold *et al.*[4,5] (see Table I). Curves are fits to a multistage model; see text for details. ×, data for rats; +, TD_{50} for rats; ◇, data for mice; □, TD_{50} for mice. The highest dose point for the vinyl chloride rat data (0.929 mg) was not used for determining the best fit curve (dashed line), since Gold *et al.* did not use this point for deriving the TD_{50}.

Gallo do not discuss these options nor do they introduce any new ones.

It is a mystery to me why Wartenberg and Gallo have chosen as "arbitrary" examples those bioassays which yielded a TD_{50} of 843 mg for vinyl chloride and 225 mg for AF-2 (furylfuramide). These are listed in the Gold *et al.*[4] compilation as experiment no. 2867 (rats, both sexes, route: inhalation, liver angiosarcoma) for vinyl chloride and no. 99 (mice, female, route: diet, forestomach squamous cell carcinoma) for AF-2. Rather than choose TD_{50} values arbitrarily, a more widely practiced approach is to take the TD_{50} at the most sensitive site for which there was a statistically significant response. Setting $p < 0.01$ as a cut-off for statistical significance, I have taken the lowest TD_{50} values in mice and in rats from Gold *et al.*'s original compilation[4] and two supplements[5]; these are given in Table I, and the data are plotted in Fig. 1.

Given the fraction of animals with tumors at each dose, the computer program MSTAGE (developed by E.A.C. Crouch) was used to calculate the maximum likelihood solutions to the parameters $a_0, a_1, \ldots, a_n$ in the multistage model formulation of the probability of tumors at a given dose d:

$$P(d) = 1 - \exp[-(a_0 + a_1 \cdot d + a_2 \cdot d^2 + a_3 \cdot d^3 + \cdots + a_n \cdot d^n)]$$

where the number of fitted parameters is set equal to the number of data points. The resultant curves are plotted in Fig. 1 for vinyl chloride and AF-2. Three of the lowest TD_{50}s were from single-dose experiments; in all three, the dose was high enough to yield >45% excess tumors. For each of these experiments the best-fit curve is drawn in Fig. 1 as a solid line. The remaining experiment (vinyl chloride in rats) included several doses, but none high enough to produce a meaningful dose–response relationship in the $[P(d) - P(0)] = 0.5$ range. Here the best-fit curve is drawn as a dashed line and is extended to the point $P(d) = 0.5$ (for the purpose of crude comparison with the data depicted in Wartenberg and Gallo's Fig. 2). The TD_{50}s calculated by Gold *et al.* are also plotted in Fig. 1. The rank order of the TD_{50}s in mice is the same as the rank order in rats, even though for rats there is some (albeit limited) low-dose data for vinyl chloride. For simplicity, I have omitted the confidence limits on the TD_{50} values (given by Gold *et al.*) and on the fractional tumor incidences, although these clearly are relevant to any meaningful quantitative comparison of dose–response behavior.

Thus, at the most sensitive site the ratio $(TD_{50,AF\text{-}2})/(TD_{50,\text{ vinyl chloride}})$ is 60.8/10.6 = 5.7 for mice and 11.4/3.69 = 3.1 for rats. This is in accord with the rank order of potencies which Wartenberg and Gallo found for tumors assessed at a less sensitive site for vinyl chloride and in the less sensitive sex for AF-2, based on the dose–response behavior at the lowest doses tested. Indeed, one rationale for using the potency at the most sensitive site is that the carcinogenic response might be less dependent on threshold phenomena such as local toxicity; therefore, in the case of genotoxic compounds, such as AF-2 and vinyl chloride, a linear response is more likely to prevail at lower doses.

Professor Ames is well aware of the likelihood that

Table I. TD_{50}s at the Most Sensitive Site in Rats and Mice[a]

Chemical	Sp.	Sex	Rte.	Site	Hist.	TD_{50}[b]	Signif.	Ref.[c]
Vinyl chloride	M	f	inh	mgl	car	10.6	$p<.0005$	No. 512, GS2
	R	both	inh	mgl	adc	3.69	$p<.008$	No. 2869, G
AF-2	M	m	eat	for	mix	60.8	$p<.005$	No. 98, G
	R	f	eat	mgl	mix	11.4	$p<.0005$	No. 103, G

[a]Abbreviations—species: M (mice), R (rats); routes: inh (inhalation), eat (diet); sites: mgl (mammary gland), for (forestomach), liv (liver); histopathology: car (carcinoma), adc (adenocarcinoma), mix (mixed sites).
[b]In mg/kg-day.
[c]G is Gold *et al.*, 1984[4]; GS2 is Gold *et al.*, 1987.[5]

testing chemicals at toxic doses results in a potentially organ-specific increase in potency relative to what might be found at lower doses, *if such experiments were performed*; indeed, he has been a leader in the exposition of this concept.[6] On the subject of the relationship between toxicity and carcinogenicity, it is unfortunate that Wartenberg and Gallo confuse the issue by writing "toxic" when they mean "carcinogenic."

Wartenberg and Gallo say that they "do not address statistical estimation issues in this note because if the index is flawed, as we contend, there is no need to estimate its values." This is absurd, because the *degree* to which an index is flawed is highly relevant to any criticism of it. If, for example, it turns out that rank order of TD_{50}s is followed at low doses only for particular classes of chemicals, or only when chemicals with minimum TD_{50}s within a certain range of one another are considered as a group, then such findings could be the basis of an interesting, useful paper.

REFERENCES

1. D. Wartenberg and M. A. Gallo, "The Fallacy of Ranking Possible Carcinogen Hazards Using the TD_{50}," *Risk Analysis* **10**, 609–613 (1990).
2. B. N. Ames, R. Magaw, and L. S. Gold, "Ranking Possible Carcinogenic Hazards," *Science* **236**, 271–280 (1987).
3. J. A. Swenberg, F. C. Richardson, J. A. Boucheron, F. H. Deal, S. A. Belinsky, M. Charbonneau, and B. G. Short, "High- to Low-Dose Extrapolation: Critical Determinants Involved in the Dose-Response of Carcinogenic Substances," *Environ. Health Perspect.* **76**, 57–63 (1987).
4. L. S. Gold, C. B. Sawyer, R. Magaw, G. M. Backman, M. de Veciana, R. Levinson, N. K. Hooper, W. R. Havender, L. Bernstein, R. Peto, M. C. Pike, and B. N. Ames, "A Carcinogenic Potency Database of the Standardized Results of Animal Bioassays," *Environ. Health Perspect.* **58**, 9–319 (1984).
5. L. S. Gold, M. de Veciana, G. M. Backman, R. Magaw, P. Lopipero, M. Smith, M. Blumenthal, R. Levinson, L. Bernstein, and B. N. Ames, "Chronological Supplement to the Carcinogenic Potency Database: Standardized Results of Animal Bioassays Published Through December 1982," *Environ. Health Perspect.* **67**, 161–200 (1986); L. S. Gold, T. H. Slone, G. M. Backman, R. Magaw, M. Da Costa, P. Lopipero, M. Blumenthal, and B. N. Ames (1987), "Second Chronological Supplement to the Carcinogenic Potency Database: Standardized Results of Animal Bioassays Published Through December 1984 and by the National Toxicology Program Through May 1986," *Environ. Health Perspect.* **74**, 237–329. (1987).
6. B. N. Ames, "Mutagenesis and Carcinogenesis: Endogenous and Exogenous Factors," *Environ. and Molec. Mutagenesis* **14** (Suppl. 16), 66–77. (1989).

Risk Analysis, Vol. 10, No. 4, 1990

Comment

Assumptions of the HERP Index[1]

David G. Hoel[2]

In his development of the HERP Index, Ames *et al.*,[1] without resorting to quantitative low-dose risk estimation, attempted to identify particular hazardous situations in order to assist in the setting of priorities. This is an admirable goal. Unfortunately, for the quantitative values given by the HERP Index to make much sense, it is necessary to adopt many of the assumptions used in low-dose quantitative risk estimation.

The HERP Index is defined as the human exposure expressed as a percentage of the carcinogenic rodent potency on a dose-per-body-weight basis. This or a per surface area metric is typically used by those carrying out quantitative risk estimates. The assumption is reasonable, and Ames has adopted it for his procedure.

Next, an assumption is needed with regard to the shape of the dose–response curve. If we assume that the cancer rate is proportional to the dose rate raised to the power k, it follows that the cancer risk is quantitatively equal to 0.5 times the HERP Index raised to the power k. This is true if in the determination of the rodent TD_{50} there were no background tumors. If, however, there was a background of 20%, say, then the coefficient would be 0.4 instead of 0.5.

Applying the HERP Index to a comparison of two chemicals makes reasonable sense only if their ranking is preserved at various environmental levels. The above calculation shows that this will be the case if the two chemicals being compared have the same dose coefficient k. If one dose–response curve for chemical A is linear (i.e., $k = 1$), and the other for chemical B is quadratic (i.e., $k = 2$), then the relative ordering is not preserved. In such circumstances, applying the HERP Index would not make much sense in that chemical A's risk would be greater than chemical B's risk, and yet the HERP Index for chemical B would be equal to that for chemical A for equal exposures and TD_{50}s. This latter problem is essentially the main argument which Wartenberg and Gallo address.

There is a second issue of the quantitative nature of the HERP Index. If one is told the HERP Index for one chemical is 0.1 and the HERP Index for another chemical is 0.01, there may be the tendency to believe that there is a similar difference in risk at the given exposure levels. This is exactly the situation if we assume that the dose–response is linear, in which case the risk is nothing more than 0.5 times the HERP Index. Therefore, Ames *et al.* is actually close to simply applying the usual low-dose linear risk estimation methods.

For example, Ames and Gold[2] estimate and compare the HERP Index for Alar from a daily glass of apple juice (0.0017%) with the HERP Index for consuming one raw mushroom per day (0.1%), a relative difference of 59. Using a linear, low-dose extrapolation, the corresponding lifetime cancer risk from drinking a daily glass of alar-contaminated apple juice is roughly 8×10^{-5}, whereas the risk from eating a daily mushroom is about 4×10^{-3}. It is difficult to imagine a situation where one is able to rank order these low-dose risks and at the same time argue that such absolute quantification of risk is unreasonable.

The HERP Index is constructed by first determining the rodent TD_{50}, based on the results from long-term carcinogenesis studies in rodents. The estimation of the TD_{50} has specific problems. In a paper by Bernstein *et al.*[3] it was shown that any estimated TD_{50} is closely related to the maximum tolerated dose (MTD) which is typically the highest dose used in the bioassay. Specifically, if one is dealing with a potent carcinogen, then it is not possible to estimate the true TD_{50} because the range of the estimated TD_{50} must be within an order of magnitude or so of the MTD. This phenomenon is a result of the discreteness of the data, and a clear example of this problem is the case of ethylene dibromide, where the tumor incidence is about 90% at all of the experimental doses in the rodent bioassay. With this type of data one does not have a reasonable estimate of where

[1] Received September 18, 1989; revised November 13, 1989.

[2] National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, North Carolina 27709.

0272-4332/90/1200-0623$06.00/1

the dose–response will turn over, and as such, one cannot provide a rational estimate of the TD_{50} (see Hoel and Bailer[4]).

The second issue is the implicit assumption of linearity of dose–response or at least the same value of k for the chemicals under consideration. It is on this area that Wartenberg and Gallo particularly concentrate. Ames *et al.* argue that a majority of carcinogens are tested at very high doses and as such general toxicity and cellular proliferation in particular are responsible for experimentally observed carcinogenesis. If this is indeed the case, then the dose–response relationships would typically be expected to be nonlinear. Wartenberg and Gallo are quite correct in expressing their concern about this general assumption, which is the key to applying the HERP Index. Much of what has been written about the HERP Index has concentrated on its application to the chemicals given in the Ames paper, and in particular the issue of natural vs. synthetic carcinogens. Not much attention, however, has been given to the technique itself, and this is what Wartenberg and Gallo address.

There is a basic problem with the indiscriminate application of any general risk formulae without a careful examination of the particular chemical in question. For example, in the Ames paper, one of the compounds with the highest HERP Index value that Ames calculates is phenobarbital. This compound is shown to have a HERP Index of 16% for a daily dose of 60 mg. Many epileptics are on maintenance doses of this drug in the range of 100–1000 mg/day, which suggests a HERP Index of about 25–250%, implying that these individuals are under extremely high risk of cancer. When researchers have actually looked at the epidemiological evidence for cancer among epileptics using phenobarbital, the picture is quite different. Ames *et al.*, in their paper, clearly point out this lack of any evidence of human cancer associated with phenobarbital (see IARC[5]), but still presented the HERP estimate. Unfortunately, the general public was shown only the table of HERP values and not the qualifying text.

The indiscriminate application of simple formulae to experimental carcinogenesis data will often be quite misleading. If one wishes to compare a glass of orange juice with a peanut butter sandwich or a mushroom, it is necessary to carefully examine all of the experimental and epidemiological data and to incorporate, as best as one can, information or conjecture about the mechanism of action before it is reasonable to make statements about either risk or diet and health.

REFERENCES

1. B. Ames, R. Magaw, and R. Gold, *Science* **236**, 271–280 (1987).
2. B. Ames and L. Gold, *Science* **244**, 755–757 (1989).
3. L. Bernstein, L. Gold, B. Ames, M. Pike, and D. Hoel, *Risk Analysis* **5**, 263–264 (1985).
4. D. Hoel and A. Bailer, "TD_{50} Estimation for Carcinogens Whose Potencies Fall Outside the Range of the Data." Unpublished manuscript.
5. International Agency for Research on Cancer (IARC) Working Group on the Evaluation of Carcinogenic Risks to Humans, "Overall Evaulations of Carcinogenicity," An Updating of IARC Monographs Volumes 1–42 (IARC Monographs, Supplement 7, 1987).

Risk Analysis, Vol. 10, No. 4, 1990

Response

A Response to Comments

D. Wartenberg[1] and M. A. Gallo[1]

1. INRODUCTION

We thank each of the discussants for providing such useful and insightful comments on our manuscript. They have helped point out limitations in our presentation and have highlighted areas of controversy. Below, we address the points we consider most salient: the model, its theoretical validity, its empirical validity, and its application to priority setting.

Models are constructs designed to summarize general data patterns and trends. They need not fit all the data observations (some even say it is the exceptions that validate the rule), but when data deviate from the valid model, the deviations should be attributable to measurement errors, stochastic variation, and the like, rather than the operation of an alternative process. Models can be used to interpolate between observations and, with caution, to extrapolate beyond the range of observation assuming homogeneity of process. Data that deviate sufficiently from one model generally inspire new models and new hypotheses. It is often the data that deviate from models that are most informative in the study of a process. Data not used in the construction of the model can be used to test process-based hypotheses, to evaluate the sensitivity of generalizations to the assumptions of a model, and to provide a theoretical framework for planning future investigations.

We argued that the TD_{50} is an inappropriate model for low-dose extrapolation of carcinogenic potency data derived from rodent bioassays. Most discussants accept our theoretical argument but claim (or show) empirically that application of the TD_{50} model to available data gives results that are consistent with results from other models applied to the same data. While this consistency among models is a useful observation, it does not validate either model, although it does provide supporting evidence for the theoretical concepts underlying each. While the TD_{50} model may be consistent with most extant data, we show analytically that it need not be consistent with all data. Empirically, we show, in fact, that it is not consistent with all currently available data. Our interpretation of these results is: (1) the TD_{50} model, if used at all for extrapolation or interpolation of data, should be used with extreme caution and scrutiny of the data for the substances under consideration; and (2) the use of the TD_{50} as a model of the carcinogenic process is problematic. Further, as we state at the conclusion of our paper, the TD_{50} model and its use in calculating a HERP convey information about only one outcome from exposure to a single chemical, ignoring issues of multiple outcomes from the same exposure,(1) chemical interactions, sensitive populations, and time history of the exposure (see Ref. 2 for a discussion of these issues). Thus, in priority setting, it is a very limited tool and requires extensive stipulation of its many assumptions.

2. THEORETICAL VALIDITY

Our initial presentation addressed the theoretical validity of the TD_{50} concept for use in low-dose extrapolation of carcinogenic potency data. Hoel, Krewski, and Gold *et al.* agree that in principle the TD_{50} model is flawed. Krewski, and Gold *et al.* argue this lack of theoretical validity is not serious because empirical results show that the model works (see below). We note, as does Hoel, that the TD_{50} model has implicit an assumption that the rank order of chemical potency is invariant under all (or all low) doses. This single assumption, we believe, is too restrictive for meaningful application of the model for potency comparison. Others (e.g., see Refs. 3–5) question the reliability of point estimates of risk generically, including models such as the TD_{50}.

Gold *et al.* argue that their HERP index, which is partially based on the TD_{50} "uses the same animal results and similar statistical methods as the usual low-dose linear estimation of risk." We agree but suggest that the TD_{50} is more restrictive: it is linear throughout all doses and requires equal dose–response curve slopes

[1] Department of Environmental and Community Medicine, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, Piscataway, New Jersey 08854–5635.

0272-4332/90/1200-0629$06.00/1

for all compounds. Further, as Bailar *et al.*[6] have shown, the assumption of low-dose linearity as conservative is not valid for all substances. Gold *et al.* counter by pointing out that using a quadratic model instead of a linear model would cause up to only a fivefold shift in the TD_{50}. We do not find this observation compelling because the TD_{50} is in the region of the observed data and our primary concern is the estimation of potency at very low doses, far from these data. Our concern with this (or any) model is the "tail wagging the dog phenomenon," that the doses used in rodent bioassay experiments do not give us sufficient information to extrapolate linearly to low doses regardless of the fit at high doses. This criticism is particularly apt for the TD_{50} model but applies to other less stringent models as well. Further, as shown by Bernstein *et al.*,[7] the range of potency estimates in a given experiment is severely limited by the number of animals used and the dosages given.

Goodman argues that the TD_{50} model is acceptable and that its limitations are well known. Yet, when we have seen the model applied either as the TD_{50} or the HERP, we have not seen any discussion of the these issues. Even in the example she cites, the discussion questions only whether the HERP overestimates the risk without considering a potential underestimate. We disagree with Goodman that use of the model is acceptable because limitations are pointed out in the literature. (Hoel points out another such situation where sufficient qualification of risks is wanting: Ames and Gold's comparison of the lifetime risk of drinking contaminated apple juice and eating mushrooms.) Rather, for extrapolation of extremely limited sample data, we question the prudence of using a model without mechanistic validity, as we will elaborate below. Further, as the results from any single stage of a complex, multistage process, such as a quantitative risk assessment, get combined into a single summary index, the limitations of each stage, most often, go unreported. We believe that most individuals who use the TD_{50} in policy evaluation are not aware of these limitations. Thus, theoretical validity is of importance.

We also acknowledge and have corrected our sloppy terminology in the draft of our original manuscript, originally using the word toxic once where we meant carcinogenic, as pointed out by Goodman.

3. EMPIRICAL VALIDITY

To bolster our theoretical argument, we selected two compounds that violate the concept behind the TD_{50} model. We did not choose the chemicals arbitrarily, as suggested by Goodman, but with careful forethought. Our goal was not to say that the model never works, but to show that in spite of the small proportion of compounds that have been subjected to rodent carcinogenicity tests and compiled by Gold *et al.*,[8–11] it is possible to find at least one pair that violate the model.

Gold *et al.* argue that our results are incorrect as we did not take account of their adjustment of the TD_{50}s to the standard lifetime. However, if data at each dose are adjusted by the same factor that Gold *et al.* use for the TD_{50}s, one would see the same effect: while AF-2 is more potent at high doses, vinyl chloride is more potent at low doses. As the TD_{50} is used to extrapolate high-dose potency to low-dose estimates, use of the estimates for comparing these two compounds below 100 mg/kg/day (adjusted for standard lifetime) would erroneously rank AF-2 as more potent than vinyl chloride (see Table I and Fig. 1). We assume that they are not suggesting adjustment only of the TD_{50} but also the experimental data. The standard lifetime adjusted data show more clearly the close correspondence between the does–response curve and the extrapolated TD_{50}s, and that the TD_{50} is a relatively good estimate of high-dose potency, for these chemicals. However, this adjustment does not resolve the issue of intersecting dose–response curves from which extrapolations to low doses can yield incorrectly ordered potencies. The point is that these two chemicals do not behave with respect to one another the way the TD_{50} model prescribes (and the way most chemicals behave with respect to one another), and other pairs of chemicals may exhibit similarly problematic patterns.

Gold *et al.*, and Goodman, argue that we did not apply the TD_{50} properly, that many other tests (e.g., target sites at which the substance is most potent, results for rats and mice) would be considered in developing a HERP. Goodman goes further by conducting an analysis of alternative data that yields results different than ours. We do not disagree with any of this, but question its relevance. Our point is that the TD_{50} model does not fit the data for these two compounds. This is not to say it does not fit any, or even most, compounds, or that various modifications of the procedure cannot be devised to fit extant data. We conclude that either the experiments we chose were errant and should not be in the database, or the model as presented and used by some investigators does not fit all the compounds. And, the model structure will not show this problem to users.

Goodman's analysis of data does not shed much light on this problem. She presents data for only one dose other than controls for three of the four experiments she discusses. Use of two data points precludes identification of any low-dose structure in the data. Ratios of TD_{50}s do not provide information about relative low-

Table I. Dose–Response Data Adjusted to Standard Lifetime[a]

AF-2			Vinyl chloride		
Dose (mg)	*P(d)*	*P(d) − P(0)*	Dose (mg)	*P(d)*	*P(d) − P(0)*
0	0/65	0.00	0	0/58	0.00
37	1/50	0.02	3.4	1/60	0.02
185	25/50	0.50	17.0	3/59	0.05
			34.0	6/60	0.10
			170.2	13/60	0.22
			409.5	13/59	0.22
225		TD_{50}	843.00		TD_{50}

[a]Dose data were adjusted as the TD_{50}s had been adjusted for standard lifetime as specified by Peto *et al.*[12] Each dose was multiplied by the square of the ratio of the length of the experiment to the standard lifetime (104 weeks). For AF-2, the experiment was 62 weeks so the dose values were multiplied by 0.36. For vinyl chloride, the experiment was 135 weeks, so the dose values were multiplied by 1.69.

dose potencies and do not accommodate potentially nonlinear dose–response relationships. Even given the data presented by Goodman, it is still theoretically possible for the associated dose–response curves to cross at low dose (if we had appropriate low-dose data), representing a change in rank of the low-dose potencies of these substances. The example does not provide sufficient evidence to counter our main thesis.

Gold *et al.* argue that AF-2 and vinyl chloride are unusual in that the variation of their TD_{50}s is extreme. Indeed, this was why we chose them, to show that the TD_{50} model works only under situations typical of the majority but not totality of chemicals so far tested. In particular, the TD_{50} model does not allow for the plateau effect seen in vinyl chloride, as noted by Gold *et al.* (In fact, for most substances, we do not have sufficient information to detect the presence of a plateau—or a supralinear dose–response curve—even if one exists. Bailar et al.[6] argue that even with current data such occurrences can be found with surprising frequency.) Because the slopes of the dose–response curves for AF-2 and vinyl chloride differ markedly, the low-dose estimates of risk based on a single dose–response model cannot be accurate for both substances—they will not fit the extant data, let alone give valid extrapolations to even lower doses. Without clear and strict guidelines delineating the situations under which the model is applicable, it is likely to be misused.

A complementary problem of differential sensitivity among different populations is presented by Prehn and Lawler.[12] They compare the responses to 3-methylcholanthrene among 10 strains of mice over 100-fold range of dose and show that the rank order of susceptibility varies as a function of dose. Again, data with this type of pattern cannot be fit with the TD_{50} model; they violate its basic assumptions.

Krewski takes the notion of empirical model assessment one step further by analyzing 572 experiments from the carcinogenic potency database[8–11] using the linearized multistage model (hereafter, LMS), and comparing slopes from this model with slopes of a line connecting the TD_{50} to the origin. About 98% of these experiments agree within a factor of 10. He concludes that, in spite of the fundamental linear limitation of the TD_{50} model, the measurable inaccuracies of low-dose potency estimates based on it are small. We are concerned about two issues on which these conclusions are based. First, implicit in Krewski's analysis is the acceptance of a single model (the LMS) as appropriate for validation of all chemicals studied. Krewski has shown only the consistency of two models both of which are based on linearity for low-dose exposures. Second, 11 of the chemicals he evaluated had vastly different slopes under the TD_{50} and LMS models (greater than a factor of 10). Extrapolation to low-dose exposures for these chemicals will result in vastly different potency estimates. The uncertainty of the estimates of effect increases as the length of the extrapolation increases, or the dose decreases.[13] While some may argue that 98% of the time is better than most statistical evaluations ($p<0.05$), this analysis only assesses the compatibility of two models, the TD_{50} and the LMS, and still misses for 11 compounds. In sum, it is only an estimate of precision rather than accuracy and the latter is also of great concern.

Krewski's discussion of 2-AAF (CPDB experiment no. 45) shows that the TD_{50} estimate of 96 mg/kg/day is problematic, far exceeding an actual dose (19.5 mg/kg/day) that produced 77% tumors. The TD_{50} markedly underestimates the carcinogenic potency of this compound. The nonlinear multistage model gives a more reasonable TD_{50}, according to Krewski, of 17.2 mg/kg/

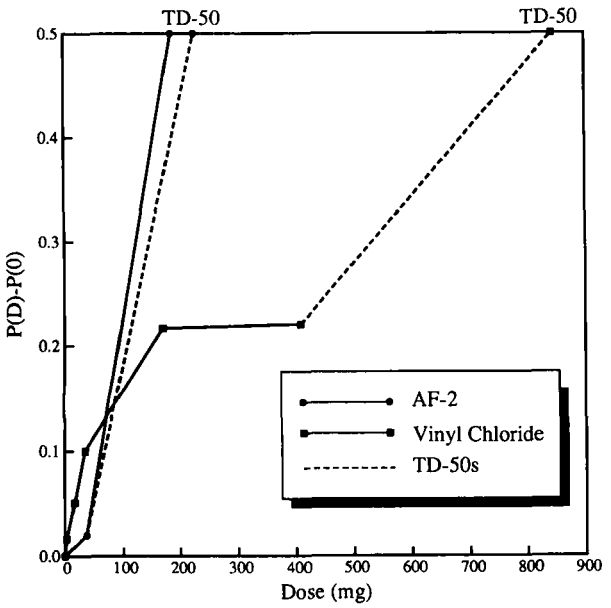

Fig. 1. Hypothetical dose-response data (solid lines) for AF-2 (circles) and vinyl chloride (squares). Data from rodent carcinogenecity experiments of different length reported in Gold *et al.*[8-11] were adjusted to a common, hypothetical, 104-week experiment duration by multiplication with the time correction factor used for adjusting the TD_{50}s to a common duration. This is done to facilitate comparison of experiments of different length in which severity of outcome is believed to be dependent on the length of the experiment. The TD_{50}s (dashed lines) shown are those reported by Gold *et al.*[8-11]. Data are provided in Table 1. Note that the TD_{50} for vinyl chloride is greater than that for AF-2 suggesting that AF-2 has greater carcinogenic potential than vinyl chloride. However, at the lower doses tested, vinyl chloride is more carcinogenic than AF-2, even after adjustment. Thus, for this one arbitrary pair of chemicals, even after adjusted to a common 104-week experiment duration, the TD_{50} concept does not correctly order cancer potencies at low dose. This same reversal is seen for other chemicals and for other dose-response models with the same chemicals (see text).

day. The disparity of these two models can be exacerbated when the data are extrapolated to low doses. They likely show a greater difference in relative risk. Use of a nonlinear model for the same chemical and the same data (and thus the same mechanism) yields a TD_{50} estimate over five times smaller (more potent), and an even greater difference in relative potency at low dose. This suggests a major problem in using the TD_{50} value for evaluation of the relative hazard of 2-AAF.

Goodman takes us to task for not addressing the issue of estimation, arguing that the degree to which the index is flawed is of importance in assessing the validity of the criticism. We agree in theory but believe the level of theoretical inappropriateness of the TD_{50} model is so great (i.e., since it cannot accomodate dose–response curves that cross each other resulting in a potency rank reversal) that estimation errors are of minor importance. We agree with her that identification of substances that violate the TD_{50} model would be extremely useful and helpful in the study of carcinogenesis, as noted above. Hoel raises additional issues regarding the estimability of the TD_{50}, particularly for potent carcinogens.

4. APPLICATION OF THE TD_{50} TO PRIORITY SETTING

Gold *et al.*'s goal of identifying hazardous situations without the need for complex, statistical models is indeed an admirable one. However, as noted by Hoel, to achieve this goal one is obliged to use a model with many of the same assumptions and limitations as these complex models.

Gold *et al.* state that their ranking, "indicates what percentage of the rodent tumorigenic dose a human gets from a given exposure." Their goal is, "to achieve some perspective on the natural background of carcinogens and to suggest priorities for epidemiological investigations." In epidemiological terms, they seek to identify substances with high attributable risk rather than high relative risk. Our concern is that the TD_{50} itself is an inappropriate model. We have not evaluated its use in the HERP. Gold *et al.* argue that with some screening and summarizing of the TD_{50} values, the HERP is useful in comparing risk. We believe, at best, that some HERP values will be misused because of inappropriate TD_{50} values.

Krewski, while agreeing with our analytic results, chides us for taking, "an unduly pessimistic view," of the use of the TD_{50}. In spite of his caveats noting the "impossibility of summarizing the characteristics of a chemical carcinogen in a single index," he favors the use of the TD_{50}. We believe that in view of the close scrutiny of quantitative risk assessment in legal and public health settings (e.g., facility siting) and the far-reaching political and policy implications of potency rankings (e.g., Alar), one must utilize methodology that is consistent and defensible for all compounds studied. To be protective of the public health, it is not sufficient to say that a model is good most of the time.

Hoel notes, as we do, that one must use more information than the simple-to-calculate HERP in assessing relative degrees of hazard if one is to be prudently protective of public health. This includes human epidemiological data and rodent bioassay data for all potential adverse outcomes (e.g., carcinogenic, neurological,

immunological, reproductive), as well as whatever mechanistic information is available.[14]

Goodman, in attacking our criticism of the TD_{50} model, claims that while Ames *et al.* "are not arguing for more widespread use of rodent potencies measured at high dose (e.g. the TD_{50}) in determining human risk," they do propose, "a simple method to combine human exposure information with rodent carcinogenicity data for obtaining an index of expected human risk (i.e., the HERP)." However, Goodman misstates our position. We do not criticize the combination of human exposure and rodent carcinogenicity data. We criticize the use of the TD_{50} model in assessing rodent carcinogenicity for use with the human-exposure data. The model is too simplistic and not sufficiently responsive to results of the rodent bioassays to provide a valid ranking of likely human response at low-dose exposure.

REFERENCES

1. H. Kipen, and D. Wartenberg, "Don't Close the Door: Some Observations on Cancer Cluster Investigation—Letter to the Editor,"*Journal of Occupational Medicine* **30**, 661–662 (1988).
2. *Complex Mixtures: Methods for In Vivo Toxicty Testing* (National Academy of Sciences, Washington, DC, 1988).
3. D. J. Paustenbach, "Important Recent Advances in the Practice of Health Risk Assessment: Implications for the 1990s," *Regulatory Toxicology and Pharmacology* **10**, 204–243 (1989).
4. R. L. Sielken, "Some Issues in the Quantitative Modeling Portion of Cancer Risk Assesment," *Regulatory Toxicology and Pharmacology* **5**, 175–181 (1985).
5. R. L. Sielken, "Quantitative Cancer Risk Assessments for TCDD," *Food Chemistry and Toxicology* **25**, 257–267 (1987).
6. J. C. Bailar III, E. A. C. Crouch, R. Shaikh, and D. Spiegelman, "One-Hit Models of Carcinogenesis: Conservative or Not?" *Risk Analysis* **8**, 485–497 (1988).
7. L. Bernstein, L. S. Gold, B. N. Ames, M. C. Pike, and D. G. Hoel, "Some Tautological Aspects of the Comparison of Carcinogenic Potency in Rats and Mice," *Fundamental and Applied Toxicology* **5**, 79–86 (1985).
8. L. S. Gold, C. B. Sawyer, R. Magaw, G. M. Backman, M. de Veciana, R. Levinson, N. K. Hooper, W. R. Havender, L. Bernstein, R. Peto, M. C. Pike, and B. N. Ames,"A Carcinogenic Potency Database of the Standardized Results of Animal Bioassays," *Environmental Health Perspectives* **58**, 9–319 (1984).
9. L. S. Gold, M. de Veciana, G. M. Backman, R. Magaw, P. Lopipero, M. Smith, M. Blumenthal, R. Levinson, L. Bernstein, and B. N. Ames, "Chronological Supplement to Carcinogenic Potency Database: Standardized Results of Animal Bioassays Published through December 1982," *Environmental Health Perspectives* **67**, 161–200 (1986).
10. L. S. Gold, T. H. Slone, G. M. Backman, R. Magaw, M. Da Costa, P. Lopipero, M. Blumenthal, and B. N. Ames, "Second Chronological Supplement to the Carcinogenic Potency Database: Standardized Results of Animal Bioassays Published through December 1984 and by the National Toxicology Program through May 1986," *Environmental Health Perspectives* **74**, 237–329 (1987).
11. L. S. Gold, T. H. Slone, G. M. Backman, S. Eisenberg, M. Da Costa, M. Wong, N. B. Manley, L. Rohrbach, and B. N. Ames, "Third Chronological Supplement to the Carcinogenic Potency Database: Standardized Results of Animal Bioassays Published through December 1986 and by the National Toxicology Program through June 1987,"*Environmental Health Perspectives* **84**, (1989).
12. L. M. Prehn, and E. M. Lalwer, "Rank Order of Sarcoma Susceptibility Among Mouse Strains Reverses with Low Concentrations of Carinogens," *Science* **204**, 309 (1979).
13. C. N. Park, "Mathematical Models in Quantitative Assessment of Risk," *Regulatory Toxicology and Pharmacology* **9**, 236–243 (1989).
14. USDA Forest Service Pacific Northwest Region, *Managing Competing and Unwanted Vegetation*. Draft Environmental Impact Statement (Portland, Oregon, October 1987).
15. R. Peto, M. C. Pike, L. Bernstein, L. S. Gold, and B. N. Ames, "The TD_{50}: A Proposed General Convention for the Numerical Description of the Carcinogenic Potency of Chemicals in Chronic-Exposure Animal Experiments," *Environmental Health Perspectives* **58**, 1–8 (1984).