To appear in *J. Reg. Toxicol. Pharmacol.*

15 April 1996

CONCORDANCE BETWEEN RATS AND MICE IN BIOASSAYS FOR CARCINOGENESIS

by

David A. Freedman
Department of Statistics
University of California
Berkeley, California 94720

Lois S. Gold
Carcinogenic Potency Data Base
University of California
Berkeley, California 94720

Tony H. Lin
Department of Statistics
University of California
Los Angeles, California 90066

Short title: Concordance in Bioassays for Carcinogenesis

Abstract

According to current policy, chemicals are evaluated for possible cancer risk to humans at low dose by testing in bioassays, where high doses of the chemical are given to rodents. Thus, risk is extrapolated from high dose in rodents to low dose in humans. The accuracy of these extrapolations is generally unverifiable, since data on humans are limited. However, it is feasible to examine the accuracy of extrapolations from mice to rats. If mice and rats are similar with respect to carcinogenesis, this provides some evidence in favor of inter-species extrapolations; conversely, if mice and rats are different, this casts doubt on the validity of extrapolations from mice to humans.

One measure of inter-species agreement is *concordance*, the percentage of chemicals that are classified the same way as to carcinogenicity in mice and rats. Observed concordance in NCI/NTP bioassays is about 75%, which may seem on the low side—because mice and rats are closely related species tested under the same experimental conditions. However, observed concordance could under-estimate true concordance, due to measurement error in the bioassays—a possibility demonstrated by Piegorsch et al. Expanding on this work, we show that the bias in observed concordance can be either positive or negative: an observed concordance of 75% can arise if the true concordance is anything between 20% and 100%. In particular, observed concordance can seriously over-estimate true concordance.

1. Introduction

According to current regulatory policy, chemicals are tested for carcinogenicity in rodent bioassays, where rats and mice are exposed to near-toxic doses of the agent on test. High doses are needed in order to demonstrate a statistically significant response with a limited number of animals. But there is an upper bound: if the dose level is set too high, animals will not live long enough to develop cancer. Thus, chemicals are administered at the "Maximum Tolerated Dose," or MTD. (Details on the MTD and bioassay design are in Section 2.)

Typically, the MTD is orders of magnitude higher than the environmental exposures of concern for the general population. To use bioassay results for risk assessment, then, two extrapolations are needed: (1) the species extrapolation from rats or mice to humans, and (2) the extrapolation from high dose to low dose. The first extrapolation is qualitative; the second is quantitative and depends on a dose-response model like the "one-hit model" (Section 2).

The focus of the present paper is the validity of the qualitative extrapolation. It is often said that most known human carcinogens are also animal carcinogens. This familiar argument, however, faces certain empirical difficulties (Freedman and Zeisel, 1988). Moreover, the argument bypasses the main question of policy interest—are most animal carcinogens also human carcinogens?

Indirect evidence can be used to validate the species extrapolation; for example, the accuracy of extrapolations from mice to rats can be examined. If mice and rats are similar with respect to carcinogenesis, this

provides some evidence in favor of inter-species extrapolation; conversely, if mice and rats are different, this casts doubt on extrapolations from rodents to humans. Data from National Cancer Institute/National Toxicology Program (NCI/NTP) are convenient for this purpose. NCI/NTP bioassays are run on a standard protocol and (with few exceptions) each chemical is tested both in rats and in mice.

Using the Carcinogenic Potency Data Base, we identified 297 chemicals tested by NCI/NTP in female mice and female rats (Gold et al., 1984, 1986, 1987, 1990, 1993). We classified each chemical as positive (+) or negative (−) in the female mouse and in the female rat, based on significance at the .005 level, one sided. This rule produces a classification in good agreement with "authors' opinion" (Haseman, 1983b; Gold et al., 1989). Being mechanical, the rule is subject to simulation study; using females avoids complications created by sex-specific responses. (Results for males are quite similar, although concordance is a bit lower.)

One measure of inter-species agreement is *concordance*, the percentage of chemicals that are classified the same way in both species. Results for NCI/NTP bioassays are shown in Table 1. There were $53+48+22+174 = 297$ chemicals; of them, $53+174 = 227$ were classified the same way in mice and in rats; the concordance is $227/297 = 76\%$. (Concordance has been computed by a number of authors, and 75% is a typical figure; see Gold et al. 1989 or Krewski et al. 1993.)

Table 1. Concordance of 297 NCI/NTP bioassays; females.

		Rats	
		+	−
Mice	+	53	48
	−	22	174

Mice and rats are, after all, very similar species being tested under virtually identical experimental conditions; it might therefore be argued that a concordance of 76% brings into question the validity of the extrapolation from rodents to humans. A possible counter-argument: the concordance observed in the NCI/NTP data is just an estimate based on limited data. Since each bioassay only involves a relatively small number of mice and rats, statistical power may be low. Thus, observed concordance could be lower than true concordance, just due to measurement error in the bioassays. Indeed, an observed concordance of 76% could imply a true concordance near 100%.

This paper follows Piegorsch et al. (1992) in exploring the question via computer simulations of the bioassay process. We expand the framework to include the case where true concordance is less than 100%, and make the simulations more realistic in other ways too. The data generated in our simulations look rather like the real NCI/NTP data with respect to

summary statistics on potency and toxicity. We show that observed concordance can be 76%—the value in Table 1—if true concordance is 20%, 100%, or anything in between. Thus, a variety of models more or less fit the data, but have quite different implications for bias in observed concordance; we doubt the data suffice to determine the bias, or give any very precise estimate of the true concordance of rats and mice.

Can risks be extrapolated from mice to rats? Previous arguments in the literature do not demonstrate the validity of the extrapolation. (Nor do we demonstrate invalidity.) The question remains open, as do more serious questions about extrapolations from rodents to humans. The statistical implication is worth stating: simulation results may be driven by assumptions rather than data.

The balance of this paper is organized as follows. Section 2 gives some detail on bioassays and the one-hit dose-response model. Section 3 describes previous simulation studies, identifies the crucial assumptions, and compares the results to real data. Section 4 describes our simulations. There is a discussion and literature review in section 5, while technical details are given in section 6. This paper is based on Lin et al. (1995), which may be consulted for additional results and further explanations, and a review of work on the quantitative extrapolation.

2. Background

Standard NCI/NTP bioassay protocols call for testing chemicals in two species (mice and rats) and in both sexes. For a given sex and species, there are three dose groups (high dose, low dose, control), each with 50 animals.

The high dose group is given the Maximum Tolerated Dose (MTD), estimated using data from a subchronic experiment; the MTD is the dose that is expected to produce a 10% decrement in weight gain but does not cause death or overt toxicity (Sontag et al., 1976). The low dose group receives half the MTD. The control group receives none of the chemical.

The probability that an animal develops cancer is often assumed to follow the one-hit model:

$$P(\text{cancer}) = p_0 + (p_{\max} - p_0)(1 - e^{-bD}). \tag{1}$$

In equation (1), p_0 is the background rate of tumors, $p_{\max}$ is the maximum probability of developing cancer, and D is the dose; $p_{\max}$ is usually taken to be 1; smaller values may be used to reflect residual genetic heterogeneity in the test animals, errors in tumor detection at necropsy, and other forms of misspecification in the conventional one-hit model. The parameter b in equation (1) is the *potency*; if a chemical is not a carcinogen, its potency is zero, by definition. The one-hit model can be fit to bioassay data to estimate the potency, as in Crouch et al. (1987) and Shlyakhter et al. (1992). The Cochran-Armitage Trend Test (Snedecor and Cochran, 1967; Gart et al., 1986) can be used to determine if bioassay results are "statistically significant," meaning they show a significant (positive) trend with dose. On genetic heterogeneity, see Gaylor et al. (1993) or Peto et al. (1985, p.46).

3. Previous Simulations

This section will summarize the simulation model used by Piegorsch et al. (1992); details are in section 6 below; also see Lin et al. (1995). The

model has three parameters: p_0, the background rate of cancer; ρ, which controls the inter-species correlation; and α, a one-sided significance level. Based on these parameters, 2000 sets of 100 "chemicals" are generated, a chemical being characterized by the quadruplet (d_m, b_m, d_r, b_r), where d is the MTD and b is potency; the subscripts m and r stand for mice and rats, respectively. In this model, by assumption, all "chemicals" are in fact carcinogenic to both species—the values for d and b are positive and finite.

Each "chemical" is subjected to a simulated NCI/NTP bioassay involving two species (mice and rats), three dose groups (control, low dose, high dose), and 50 animals per dose group. The probability of cancer follows the standard one-hit model, equation (1) with $p_{\max} = 1.0$. A chemical is classified as "+" if a Cochran-Armitage Test on the bioassay results shows a statistically significant positive trend at the α level, one-sided. This leads to a classification as "++", "+−", "−+", or "−−", where the first and second symbols denote the results in mice and rats, respectively.

For each set of 100 chemicals, the concordance is computed. Then, the 2000 concordances are averaged. This entire process is repeated for many different values of p_0, ρ, and α. The principal finding is that observed concordances are less than true concordance, with an upper bound of about 80%.

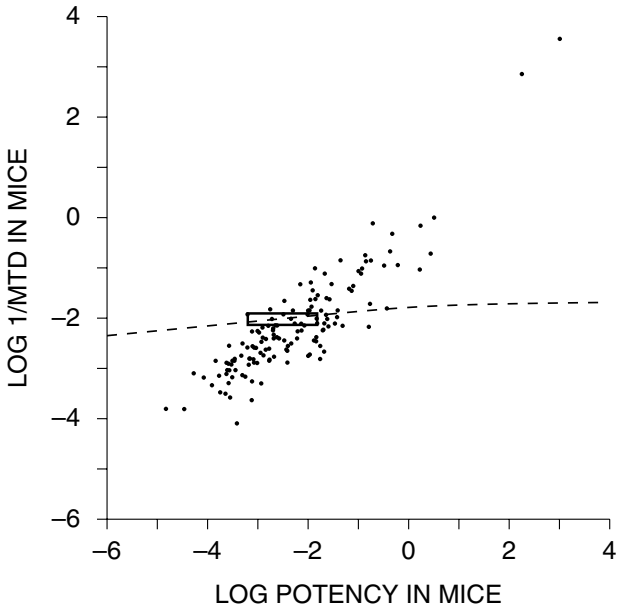
Piegorsch et al. report that $p_0 = .10$, $\rho = .9$, and $\alpha = .025$ give simulated concordances that are similar to NCI/NTP data (Table 1). However, other aspects of their simulation are quite unrealistic, as shown in Figure 1 for mice (the plot for rats would be similar). The horizontal axis shows $\log_{10}$potency; the vertical axis shows $\log_{10}(1/\text{MTD})$. Each of the 143 dots

corresponds to an NCI/NTP bioassay that had significant results in mice at the .025 level. The dashed line is the graph of equation (4) below, which is the theoretical relationship between toxicity and potency built into the simulation model. The real NCI/NTP data do not follow the theoretical line.

We computed the box in Figure 1 by generating 100,000 statistically significant ($\alpha = .025$) chemicals according to the procedure described above, with $p_0 = .10$ and $\rho = .9$. The horizontal edges of the box show the mean log potency, plus or minus three standard deviations. The vertical edges of the box show the mean $\log_{10}(1/\text{MTD})$, plus or minus three standard deviations. Among the 100,000 simulated chemicals, 98.1% had values inside the box. By contrast, among the 143 NCI/NTP chemicals, only 8 had values inside the box. The box covers only a very small part of the real data: you may need to look closely at the figure to see the box.

There is another unrealistic assumption that drives the results in Piegorsch et al.'s simulations: all chemicals are assumed carcinogenic both for mice and for rats—so true concordance is 100%. It is therefore not surprising that concordance is under-estimated. The observed concordance has nowhere to go but down.

Figure 1. The simulation in Piegorsch et al. (1992) compared to NCI/NTP data; chemicals that are statistically significant carcinogens, .025 level; logs to base 10; female mice.



4. New Simulations

This section presents results from new simulations with more plausible assumptions. Following Piegorsch et al., we generate chemicals for testing as a random sample from some (hypothetical) population of chemicals, whose true potencies are unobservable. After a chemical is selected, it is run through a simulated bioassay, just as in Piegorsch et al. The bioassay provides an estimated potency, which may differ from the true potency, because there are only a finite number of animals on test. The bioassay also classifies each chemical as a carcinogen or a non-carcinogen in each species; the observed classification may differ from the true classification, due to statistical error. We define the population so that some chemicals are in truth

non-carcinogenic. The true (unobservable) potencies and MTDs are chosen so the distribution of estimated potencies and MTDs looks rather like the NCI/NTP data. In particular, with our simulations, observed concordance will be about 76%. However, the true concordance—in the population of all possible chemicals—ranges from 20% to 100%. These results suggest that bias in observed concordance is not determined by the data.

Table 2. Simulation results: four models. Percentages based on samples of size 297.

	True Classification				Concordance	
Model	++	+−	−+	−−	True	Observed
A	20	20	30	30	50	76
B	18	53	29	0	18	76
C	20	5	5	70	90	92
D	47	0	0	53	100	76

Table 2 shows four variations on this theme. (For details, see section 6 below, or Lin et al., 1995). For each line of the table, we drew 1000 samples; each sample had 297 chemicals, as in the NCI/NTP data. Different lines in the table are based on different theoretical populations of chemicals. For example, take model A (line 1). As shown by the first 4 entries, 20% of the chemicals in the population are carcinogenic in mice and rats; 20% are

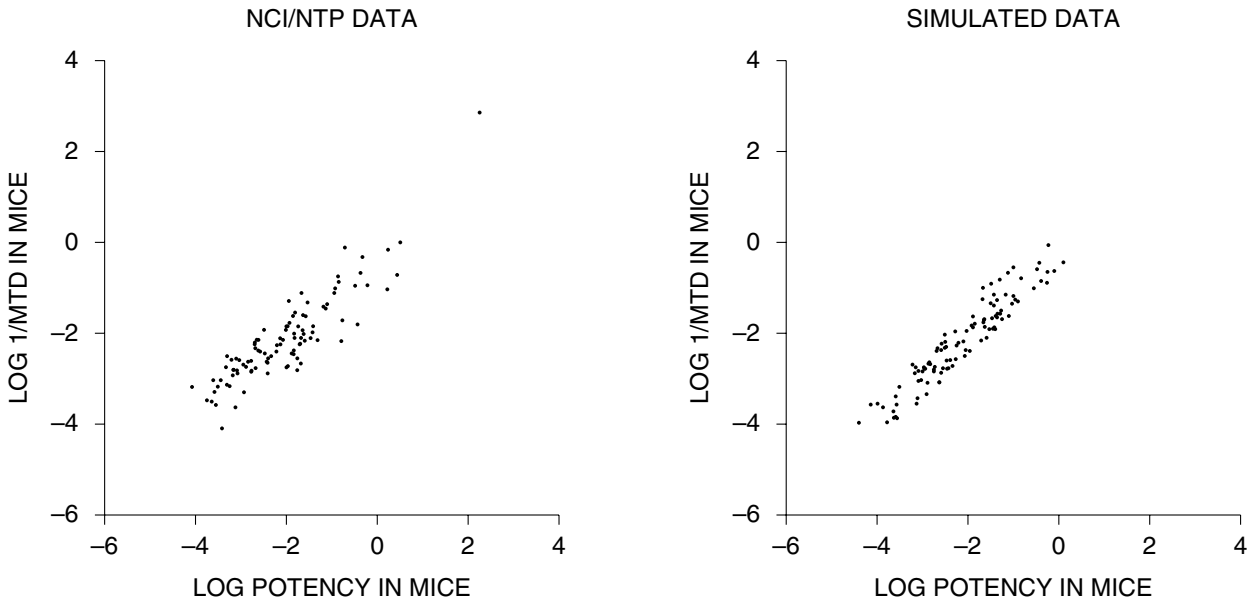
carcinogenic in mice but not in rats; 30% are non-carcinogenic in mice but carcinogenic in rats; 30% are not carcinogenic to either species. This is—by construction—truth in model A. Thus, true concordance is (5th entry)

$$20\% + 30\% = 50\%.$$

In this model, carcinogenicity is independent between the two species: the mouse carcinogen, just like the mouse non-carcinogen, has a 50% chance to be a rat carcinogen. Information about carcinogenicity in one species gives no information about carcinogenicity in the other species. Yet observed concordance is 76%, just as in the NCI/NTP data. Figure 2 compares the simulated bioassay data in mice for this model to the real NCI/NTP data; the match is reasonable—but hardly perfect. Similar conclusions would hold for rats, for mice vs. rats, etc.

For model B (line 2 of Table 2), true concordance is 18% and there is an inverse relationship between the two species. In truth, among the mouse carcinogens, only $18/(18 + 53) = 25\%$ are rat carcinogens; however, among the mouse non-carcinogens, 100% are rat carcinogens. Thus, carcinogenicity in one species points in truth to non-carcinogenicity in the other. Yet observed concordance is 76%. Summary statistics on potency and toxicity also match the real data.

Figure 2. Simulated data compared to NCI/NTP data; chemicals that are statistically significant carcinogens, at the .005 level; logs to base 10; female mice.



In model C, true concordance is 90% while observed concordance is 92%, showing that current bioassay design does allow high observed concordance (for certain populations of chemicals). With this model, the observed concordance table does not—indeed, cannot—resemble the true concordance table. In other respects, however, the data seem quite realistic. For model D, true concordance is 100% but observed concordance is 76%, so that observed concordance may under-estimate true concordance, as suggested by Piegorsch et al.

What is the source of the bias?

Each chemical belongs to one of four categories, depending on "true" mouse- and rat-carcinogenicity (i.e., "++", "+−", and so forth); also, each chemical belongs to one of four categories, depending on "observed" carcinogenicity. This gives rise to a 4 × 4 matrix. Results for our first model are presented in Table 3. The column totals give the average "true" number of each type of chemical. The row totals give the average "observed" number of each type of chemical, the basis for the observed concordance in the 6th column of Table 2.

Table 3. Detailed results for the first model.

	True					
Observed	++	+−	−+	−−	Total	Percent
++	52.5	.2	.1	.0	52.8	17.8
+−	3.9	43.7	.3	.4	48.4	16.3
−+	2.7	.1	18.9	.4	22.1	7.4
−−	.2	15.4	70.0	88.1	173.8	58.5
Total	59.3	59.4	89.4	89.0	297.0	100.0
Percent	20.0	20.0	30.1	30.0	100.0	—

Recall that observed concordance is obtained by averaging 1000 sets of 297 simulated chemicals. The last row and column in Table 3 give percent-

ages, with 297 as the base. (Detail may not add to total, due to independent rounding; the last row in Table 3 differs from the first row in Table 2, due to sampling error; from the last column in Table 3, the observed concordance in model A is $17.8\% + 58.5\% = 76.3\%$; this is rounded to 76% in Table 2.)

On the average, 59.3 out of the 297 chemicals were true "++". Most of these (52.5) were observed as "++" in the simulated bioassays, but an average of $3.9 + 2.7$ were misclassified as discordant ("+−" or "−+"). Also, 89.0 chemicals were true "−−"; of these, an average of $.4 + .4$ were misclassified as discordant. The average total number of false discordances can thus be computed as $3.9 + 2.7 + .4 + .4 = 7.4$. On the other hand, the average total number of false concordances is $.2 + .1 + 15.4 + 70.0 = 85.7$. The number of false concordances is much larger than the number of false discordances: in particular, the observed "−−" cell is inflated, due to lack of power in the bioassay. This is what makes the observed concordance much larger than the true concordance.

5. Discussion

Piegorsch et al. (1992) suggest that true concordance is greater than observed concordance, especially for chemicals that are only weakly carcinogenic; indeed, an observed concordance of 75% may imply a true concordance of nearly 100%, and observed concordance may have an upper bound of about 80%:

> "investigations, using computer simulations, illustrate that the concordance underestimation can be rather severe even when restricted to a narrow range of relatively low underlying potencies. At these levels, average observed concordance may be limited to only about 80%, suggesting that observed values at or near 75% may in fact be indicative of greater agreement than previously considered ... concordance information at relatively low levels of potency can be seriously underestimated, weakening the overall measure of agreement exhibited by the data, and leading to suspect or unsure inferences. [p.119]"

These results have been cited as showing that observed concordance is biased downward, so that 80% is an upper bound on observable concordance; see, for instance, (Huff et al., 1991) and (Haseman and Seilkop, 1992). However, the results are based on assumptions about the true (unobservable) parameters governing chemical carcinogenicity. These assumptions are somewhat unrealistic (Figure 1). Furthermore, Piegorsch et al. have in effect assumed that all chemicals are carcinogenic in both species, so true concordance is 100%. On that basis, observed concordance has nowhere to go but down.

As Table 2 demonstrates, it is possible to have low true concordance but moderately high observed concordance. It is even possible to have a high true concordance and a higher observed concordance. In these models, observed concordance is biased high, on the average across all chemicals. Of course, it is also possible to have a true concordance of 100% but only moderately high observed concordance (model D).

Piegorsch et al. pointed out that bias in concordance could depend on toxicity; if so, stratification by the MTD would help. We examined this idea by computing concordance separately for chemicals with mouse MTDs above and below 100, in model A. (The units of dose are "milligrams per kilogram of body-weight per day.") As it turned out, observed concordance was higher than true concordance for both groups of chemicals, by about 25 percentage points. Stratification does not seem to resolve the problem.

So far, we have shown that a variety of models—with quite different true concordances—are more or less consistent with the NCI/NTP data. It therefore seems unlikely that the true concordance can be estimated with any reasonable degree of confidence from bioassay data, without imposing further constraints.

Like previous authors, we used a variant of the one-hit model. We made some allowance for specification error, because—if examined in detail—the one-hit model may be rejected. For reviews, see Food Safety Council (1980), Freedman and Zeisel (1988); also see Peto et al. (1984), *Cancer Research* (1991) Vol. 51 No. 23 Part 2 pp.6407–6491, Meier et al. (1993), Hoel and Portier (1994).

Too, there are familiar difficulties in using the data to discriminate among models; for a recent discussion, see Kopp-Schneider and Portier (1991). In some respects, the "multistage model" extends the one-hit model, taking into account duration as well as level of dose and time to tumor; even this more general model will not fit a number of data sets (Freedman and Navidi, 1989, 1990). Also see Moolgavkar (1990, 1994), who discusses alternative models. Because of uncertainties about dose-response models, sim-

ulation studies are rather idealized versions of reality. Such studies cannot give definitive evidence about concordance, but can indicate the complexities in estimating measures of inter-species agreement from bioassay data.

Worst-case analysis

In a bioassay, some 35 target organs are examined, and risk assessment is based on the most sensitive site. In other words, classification of carcinogenicity is based on the response at the most sensitive site, and extrapolations from rodent to human are based on the potency at this site. However, rodent carcinogens often increase the tumor rate at some sites but decrease the rate at other sites—even in the same sex-species group in the same experiment. (A further complication: animals in the treatment groups tend to weigh less, and lower body weight is associated with a reduction in tumor incidence.) We think that both the positive and the negative trends should be considered when assessing carcinogenicity—a topic not addressed by our simulations. In effect, like previous authors, we studied concordance of worst-case analyses in mice and rats. For reviews, see Haseman (1983a), Salsburg (1983), Freedman and Zeisel (1988), Davies and Monro (1994), Haseman and Johnson (1996).

6. Technical details

Piegorsch et al. (1992) use a simulation study to examine potential bias in observed concordance. The study is keyed to data from the Carcinogenic Potency Data Base of Gold et al. (1984, 1986, 1987). From this database, Piegorsch et al. select the 405 chemicals with results both in mice and in

rats. Each chemical is characterized by six numbers: d_m, the MTD in mice; b_m, the estimated potency in mice; c_m, the "carcinogenicity" in mice ("+" for carcinogens, "−" for noncarcinogens); and d_r, b_r, and c_r, for rats. If c_m is "−", then b_m is set to zero; likewise for c_r and b_r. The study uses a new measure of carcinogenicity for mice:

$$\theta_m = \ln\left(1 + \frac{b_m}{\ln 2}\right). \tag{2}$$

A similar equation defines θ_r for rats. Finally, pairs (d, θ) are obtained by pooling data for mice and rats. (Piegorsch et al. use "the literature" as well as NCI/NTP, and take the site with highest estimated potency in males or females; see their Appendix A.)

Piegorsch et al. report a regression of $\ln d$ on $\ln \theta$:

$$\ln d = 4.103 - 0.097 \ln \theta; \tag{3}$$

presumably, θ is truncated below at a small positive value. Substituting equation (2) into equation (3) yields

$$\ln d = 4.103 - 0.097 \ln\left[\ln\left(1 + \frac{b}{\ln 2}\right)\right], \tag{4}$$

where d is the MTD and b is the potency.

Each simulation is characterized by three parameters: p_0, the background rate of cancer; ρ, a parameter that controls the inter-species correlation; and α, a one-sided significance level. Based on these parameters, 2000 sets of 100 "chemicals" are generated, each "chemical" being generated as follows. Choose a pair (z_m, z_r) from a bivariate normal distribution with mean 0, variance 1, and correlation ρ; let $\theta_m = 10^{-4+2\Phi(z_m)}$ and $\theta_r =$

$10^{-4+2\Phi(z_r)}$, where Φ is the standard normal distribution function; compute the simulated MTD in mice d_m from θ_m, using equation (3); compute the simulated potency in mice b_m from the identity $b_m = (e^{\theta_m} - 1) \times \ln 2$; for rats, compute the MTD d_r and the potency b_r from θ_r. The resulting quadruplet (d_m, b_m, d_r, b_r) characterizes a simulated chemical. By construction, all simulated chemicals are carcinogenic in both species, with positive values for θ_m and θ_r. The original carcinogenicity indicators c_m and c_r and the initial measures θ_m and θ_r of carcinogenicity play no role in these simulations, except to derive equations (3) and (4).

As previously noted, each "chemical" is subjected to a simulated NCI/ NTP bioassay involving two species (mice and rats), three dose groups (control, low dose, high dose), and 50 animals per dose group. The probability of cancer follows the standard one-hit model: equation (1) with $p_{\max} = 1.0$. A chemical is classified as "+" if a Cochran-Armitage Test on the bioassay results shows a statistically significant positive trend at the α level, one-sided. This leads to a classification as "++", "+−", "−+", or "−−", where the first and second symbols denote the observed carcinogenicity in mice and rats, respectively. The test for trend is applied to tumor rates in the three dose groups; time-to-tumor is not considered.

We turn now to our simulations. Each "chemical" is generated as a set of "true" values (c_m, c_r, x_m, x_r, y_m, y_r). The values c_m and c_r indicate carcinogenicity: $c_m = 1$ for mouse carcinogens, and $c_m = 0$ otherwise; likewise for c_r. These c's are the "true" carcinogenicity indicators. The values x_m and x_r are the "true" log MTD's for mice and rats. The values y_m and y_r are the "true" log potencies for mice and rats; logs are to base 10.

For mouse noncarcinogens, $y_m = -\infty$; for rat noncarcinogens, $y_r = -\infty$. For the parameters and their rationale, see Lin et al. (1995).

In our simulations, the probability of cancer is assumed to follow the one-hit model (1), with a background cancer rate of $p_0 = 10\%$ and an upper bound of $p_{\max} = 90\%$. If $y_m = -\infty$ or $y_r = -\infty$, the corresponding probability of cancer is simply the background rate. In effect, this procedure fits the standard one-hit model ($p_{\max} = 1$) to the data, although the true value for $p_{\max}$ is 0.9. This amount of specification error does not seem unrealistic.

Each "chemical" is subjected to the simulated NCI/NTP bioassay and Cochran-Armitage Trend Tests, as described above. The bioassay and the tests generate set of "observed" values ($\hat{c}_m$, $\hat{c}_r$, x_m, x_r, $\hat{y}_m$, $\hat{y}_r$) for each chemical. The values $\hat{c}_m$ and $\hat{c}_r$ indicate statistical significance: $\hat{c}_m = 1$ if the trend for mice is statistically significant at the .005 level, $\hat{c}_m = 0$ otherwise; similarly for $\hat{c}_r$. The $\hat{c}$'s provide the "observed" classification as to carcinogenicity in the two species. The x_m and x_r are the log MTD's, observed without error. Finally, $\hat{y}_m$ and $\hat{y}_r$ are the maximum likelihood estimates for log potency, based on the bioassay data.

The procedure for generating "chemicals" is a bit complicated. The vectors of "true values" (c_m, c_r, x_m, x_r, y_m, y_r) are generated as independent and identically distributed observations from a random vector

$$(C_m, C_r, X_m, X_r, Y_m, Y_r, \epsilon_m, \epsilon_r).$$

Conditioned on C_m and C_r, the log MTD variables X_m and X_r have a bivariate normal distribution with $\mathrm{corr}(X_m, X_r) = .93$. (In the NCI/NTP

data, the correlation between X_m and X_r was .93 for the 53 "++" chemicals, and did not vary much from cell to cell in the 2×2 table.) Given C_m and C_r, the variables ϵ_m and ϵ_r are independent of each other and of the pair (X_m, X_r). If $C_m = 1$, then ϵ_m is normally distributed, and otherwise $\epsilon_m = -\infty$ with probability one; likewise for C_r and ϵ_r. Finally, the log potency variables Y_m and Y_r are defined by the equations $Y_m = -X_m + \epsilon_m$ and $Y_r = -X_r + \epsilon_r$.

Each model is completely specified by the joint distribution of (C_m, C_r, X_m, X_r, Y_m, Y_r, ϵ_m, ϵ_r); for details, see Lin et al. (1995). The statistical power of a simulated bioassay is determined by the ϵ's. Indeed, ϵ_m and ϵ_r govern tumor yield via the one hit model (1): $bD = \exp(\epsilon)$ when D is the MTD, while $bD = 0.5 \times \exp(\epsilon)$ when D is 0.5×MTD. Moreover, if a chemical is not a carcinogen, it does not cause cancer at any dose; thus, $b = 0$, $bD = 0$, $Y = -\infty$, and $\epsilon = -\infty$. In the simulations, we use the .005 level, one-sided, as noted above; changing levels from .005 to .025 would not alter the concordances appreciably; however, the 2×2 table would no longer match the NCI/NTP data so well, unless other parameters were also changed.

Authors' Footnote

We thank D. Gaylor and W. Piegorsch for helpful suggestions. Research of LSG supported by Director, Office of Energy Research, Office of Health and Environmental Research of the U. S. Department of Energy under Contract DE-AC03-76SFO0098, and through the University of Cali-

fornia, Berkeley, by the National Institute of Environmental Health Sciences Center Grant ESO1896.

References

Davies, T. S., and Monro, A. 1994. The rodent carcinogenicity bioassay produces a similar frequency of tumor increases and decreases: implications for risk assessment. *Regulatory Toxicology and Pharmacology* 20: 281–301.

Food Safety Council. 1980. Quantitative risk assessment: Report of the Scientific Committee. *Food and Cosmetics Toxicology* 18: 711–734.

Freedman, D., Gold, L. S., and Slone, T. 1993. How tautologous are interspecies correlations of carcinogenic potencies? *Risk Analysis* 13: 265–272.

Freedman, D. A., and Navidi, W. C. 1990. Ex-smokers and multistage model for lung cancer. *Epidemiology* 1: 21–29.

Freedman, D. A., and Navidi, W. C. 1989. Multistage models for carcinogenesis. *Environmental Health Perspectives* 81: 169–188.

Freedman, D., and Zeisel, H. 1988. From mouse to man: the quantitative assessment of cancer risks. *Statistical Science* 3: 3–56, with discussion.

Gart, J., Krewski, D., Lee, P., Tarone, R., and Wahrendorf, J. 1986. *Statistical methods in cancer research. Volume III. The design and analysis of long-term animal experiments.* International Agency for Research on Cancer, Lyons, France. Scientific Publication No. 79.

Gaylor, D., Chen, J., and Sheehan, D. 1993. Uncertainty in cancer risk estimates. *Risk Analysis* 13: 149–154.

Gold, L. S., Bernstein, L., Magaw, R., and Slone, T. 1989. Interspecies extrapolation in carcinogenesis: prediction between rats and mice. *Environmental Health Perspectives* 81: 211–219.

Gold, L. S., Manley, N., and Ames, B. N. 1992. Extrapolation of carcinogenicity between species: qualitative and quantitative factors. *Risk Analysis* 12: 579–588.

Gold, L. S., Manley, N., Slone, T., Garfinkel, G., Rohrbach, L., and Ames, B. N. 1993. The fifth plot of the carcinogenic potency database: results of animal bioassays published in the general literature through 1988, by the National Toxicology Program though 1989. *Environmental Health Perspectives* 100: 65–135.

Gold, L. S., Sawyer, C., Magaw, R., Backman, G., de Veciana, M., Levinson, R., Hooper, N., Havender, W., Bernstein, L., Peto, R., Pike, M., and Ames, B. N. 1984. A carcinogenic potency database of the standardized results of animal bioassays. *Environmental Health Perspectives* 58: 9–319.

Gold, L. S., Slone, T., Backman, G., Eisenberg, S., Da Costa, M., Wong, M., Manley, N., Rohrback, L., and Ames, B. N. 1990. Third chronological supplement to the carcinogenic potency database: standardized results of animal bioassays published through December 1986, by the National Toxicology Program through June 1987. *Environmental Health Perspectives* 84: 215–286.

Gold, L. S., Slone, T., Backman, G., Magaw, R., Da Costa, M., Lopipero, P., Blumenthal, M., and Ames, B. N. 1987. Second chronological supplement to the carcinogenic potency database: standardized results of animal

bioassays published through December 1984, by the National Toxicology Program through May 1986. *Environmental Health Perspectives* 74: 237–239.

Gold, L. S., Slone, T., Stern, B., and Bernstein, L. 1993. Comparison of target organs of carcinogenicity for mutagenic and non-mutagenic chemicals. *Mutation Research* 286: 75–100.

Gold, L. S., de Veciana, M., Backman, G., Magaw, R., Lopipero, P., Smith, M., Blumenthal, M., Levinson, R., Gerson, J., Bernstein, L., and Ames, B. N. 1986. Chronological supplement to the carcinogenic potency database: standardized results of animal bioassays published through December 1982. *Environmental Health Perspectives* 67: 161–200.

Haseman, J. K. 1983a. Patterns of tumor incidence in two-year cancer bioassay feeding studies in Fischer 344 rats. *Fundamental and Applied Toxicology* 3: 1–9.

Haseman, J. K. 1983b. Issues: a re-examination of false-positive rates for carcinogenesis studies. *Fundamental and Applied Toxicology* 3: 334–339.

Haseman, J. K., and Johnson, F. M. 1996. Analysis of National Toxicology Program rodent bioassays for anticarcinogenic effects. *Mutation Research* 350: 131–141.

Haseman, J. K., and Lockhart, A. 1993. Correlations between chemically related site-specific carcinogenic effects in long-terms studies in rats and mice. *Environmental Health Perspectives* 101: 50–55

Haseman, J., and Seilkop, S. 1992. An examination of the association between maximum tolerated dose and carcinogenicity in 326 long-term studies in rats and mice. *Fundamental and Applied Toxicology* 19: 207–213.

Hoel, D. G., and Portier, C. J. 1994. Nonlinearity of dose-response functions for carcinogenicity. *Environmental Health Perspectives* 102 Suppl 1: 109–13.

Huff, J., Cirvello, J., Haseman, J., and Bucher, J. 1991. Chemicals associated with site-specific neoplasia in 1394 long-term carcinogenesis experiments in laboratory rodents. *Environmental Health Perspectives* 93: 247–270.

Kopp-Schneider, A., and Portier C. J. 1991. Distinguishing between models of carcinogenesis: the role of clonal expansion. *Fundamental and Applied Toxicology* 17: 601–13.

Krewski, D., Gaylor, D., Soms, A., and Szyszkowicz, M. 1993. An overview of the report: correlation between carcinogenic potency and the maximum tolerated dose: implications for risk assessment. *Risk Analysis* 13: 383–398.

Lin, T. H., Gold, L. S., and Freedman, D. A. 1995. Carcinogenicity tests and inter-species concordance. Technical Report No. 439, Department of Statistics, University of California, Berkeley 94720. To appear in *Statistical Science*, November, 1995, Vol. 10, No. 4.

Meier, K. L., Bailar, J., and Portier, C. J. 1993. A measure of tumorigenic potency incorporating dose-response shape. *Biometrics* 49: 917–926.

Moolgavkar, S. H. 1990. Cancer models. *Epidemiology* 1: 419–20.

Moolgavkar, S. H. 1994. Biological models of carcinogenesis and quantitative cancer risk assessment. *Risk Analysis* 14: 879–82.

Peto, R., Parish, S., and Gray, R. 1985. There is no such thing as ageing, and cancer is not related to it. In *Age-Related Factors in Carcinogenesis.* (A. Likhachev, V. Anisimov, and R. Montesano, Eds.) International Agency for Research on Cancer, Lyons, France. Scientific Publication No. 58, pp. 43-54. See especially p.46.

Piegorsch, W., Carr, G., Portier, C., and Hoel, D. 1992. Concordance of carcinogenic response between rodent species: potency dependence and potential underestimation. *Risk Analysis* 12: 115–121.

Salsburg, D. S. 1983. The lifetime feeding study in mice and rats—an examination of its validity as a bioassay for human carcinogenesis. *Fundamentals of Applied Toxicology* 3: 63–67.

Snedecor, G., and Cochran, W. 1967. *Statistical methods.* Iowa State University Press, Ames.

Sontag, J., Page, N., and Saffiotti, U. 1976. Guidelines for carcinogen bioassays in small rodents. Carcinogenesis technical report no. 1. National Cancer Institute, Bethesda, MD.