Bias in Qualitative Measures of Concordance for Rodent Carcinogenicity Tests

Tony Lin
Adjunct Assistant Professor,
Department of Mathematics,
University of California, Los Angeles 90066

Lois Swirsky Gold
Director, Carcinogenic Potency Project,
Lawrence Berkeley Laboratory,
University of California, Berkeley 94720

David Freedman
Professor, Department of Statistics,
University of California, Berkeley 94720

Revised 8/25/95
Technical Report No. 439,
Department of Statistics,
University of California, Berkeley 94720

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 around 75%, which may seem on the low side—because mice and rats are closely related species tested under the same experimental conditions. Theoretically, observed concordance could under-estimate true concordance, due to measurement error in the bioassays. Thus, bias in concordance is of policy interest. Expanding on previous work by Piegorsch et al. (1992), 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 overestimate true concordance.

A variety of models more or less fit the data, with quite different implications for bias. Therefore, given our present state of knowledge, it seems unlikely that true concordance can be determined from bioassay data.

1. Introduction

According to current regulatory policy, chemicals are tested for carcinogenicity; screening is routinely done in animal experiments—*bioassays.* In a bioassay, 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). In brief, if $P(\text{cancer})$ is the probability of developing cancer at dose D, the one-hit model says

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

The model has two parameters, p_0 and b. The parameter p_0 is the background rate of cancer when the dose D is 0. The parameter b is called "potency." This parameter can be estimated from bioassay data and a chemical can be classified as carcinogenic if its estimated potency is statistically significant—in other words, $\hat{b}$ exceeds zero by an amount that is statistically significant.

The focus of the present paper is the validity of the qualitative extrapolation (although the quantitative extrapolation and the one-hit model will be discussed too). Little direct evidence is available on the qualitative extrapolation because so few chemicals have been evaluated in human studies. 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 a question of considerable 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 on rats and on 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; Gold and Manley et al., 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.

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; other literature is reviewed below.)

Mice and rats are, after all, very similar species being tested under virtually identical experimental conditions; it might be argued that a concordance of 75% is on

NCI/NTP

		Rats	
		+	−
Mice	+	53	48
	−	22	174

Table 1: Concordance table for 297 NCI/NTP bioassays

the low side, bringing 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. Theoretically, observed concordance could be lower than true concordance, due to measurement error in the bioassays; indeed, an observed concordance of 75% could imply a true concordance near 100%.

Here, we follow Piegorsch et al. (1992) in exploring this question via computer simulations of the bioassay process. We expand the framework used by those authors to include the case where true concordance is less than 100%, and we 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 75% if true concordance is 20%, 100%, or anything in between—depending on the choice of parameters. In other words, a variety of models more or less fit the data, but have radically different implications for bias in observed concordance. Thus, we doubt the data suffice to determine the bias, or give any very precise estimate of the true concordance of rats and mice—nor yet the validity of the species extrapolation from rodents to humans.

We turn to the quantitative extrapolation and inter-species correlations of carcinogenic potency. Using NCI/NTP data, Crouch and Wilson (1979) found a strong correlation between estimated potencies in rats and mice. However, Bernstein et al. (1985) showed this correlation to be explicable in terms of statistical artifact. The correlation is due to (1) the choice of data set, namely, all chemicals with potency estimates that were statistically significant in both species, (2) the high correlation between the MTDs in mice and rats, and (3) absence of 100% cancer rates in the NCI/NTP data. This work will be reviewed in section 6.

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 implications are worth stating explicitly: (1) simulation results may be driven by assumptions rather than data, and (2) correlations may be driven by selection of samples. When it comes to policy analysis, such possibilities should be carefully considered.

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. Section 5 extends the results to other measures of qualitative agreement such as the odds ratio. Section 6 discusses the quantitative extrapolation. Literature is reviewed in sections 5 and 6.

2. Background

In bioassays, animals are exposed to chemicals in order to determine carcinogenicity. Standard NCI/NTP protocols call for testing a chemical 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 preliminary experiment; the MTD is the dose that produces a 10% decrement in predicted 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. For a detailed description of bioassay design, see (Freedman and Zeisel, 1988).

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 of $p_{\max}$ may be used to reflect residual genetic heterogeneity in the test animals, errors in tumor detection at necropsy, and other forms of miss-specification in the conventional one-hit model. The parameter b in equation (1) is the *potency*; if a chemical is a 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). This model is often used, despite a number of difficulties (Freedman and Zeisel, 1988). 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 heterogeneity, see Gaylor et al. (1993), Peto et al. (1985, p.46); also see Peto et al. (1975), Peto et al. (1984).

The data in this paper cover 297 chemicals tested by NCI/NTP with results in female mice and female rats (Gold et al., 1984, 1986, 1987, 1990; Gold and Manley et al., 1993). Potencies were standardized to a two-year lifespan.

3. Previous Simulations

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 mouse carcinogens, "−" for mouse 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 $\ln \theta$:

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

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. A "chemical" is 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.

Each "chemical" is then 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 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). By construction, all simulated chemicals are carcinogenic in both species, with positive values for θ_m and θ_r chosen as described above. (The test for trend is applied to tumor rates in the three dose groups; time-to-tumor is not considered: in the jargon of the field, the analysis is based on *summary data* rather than *lifetable data.*)

For a given triple of parameters (p_0, ρ, α), 2000 sets of 100 chemicals are generated and classified. 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 the observed concordances were always less than the 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 that simulation are quite unrealistic, as shown in Figure 1 for mice (the plot for rats would be similar). The horizontal axis shows log potency; the vertical axis shows log(1/MTD); logs are to base 10. Each of the 143 dots corresponds to an NCI/NTP bioassay that had significant results in mice at the .025 level. The dotted line is the graph of equation (4), which is the relationship between MTD and potency built into the simulations. The real NCI/NTP data do not follow the theoretical line.

The box in Figure 1 was computed by generating 100,000 statistically significant ($\alpha = .025$) chemicals according to the procedure described above, using $p_0 = .10$ 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(1/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. Adding points to represent experiments in "the literature" other than NCI/NTP only accentuates the discrepancy: Piegorsch et al.'s trend line does not follow the data. For further discussion, see (Lin, 1994).

There is another unrealistic assumption that drives the results. In the simulations, all chemicals are carcinogenic both for mice and for rats by construction, so the true concordance is 100%—by assumption. It is not surprising that concordance is underestimated: the observed concordance has nowhere to go but down.

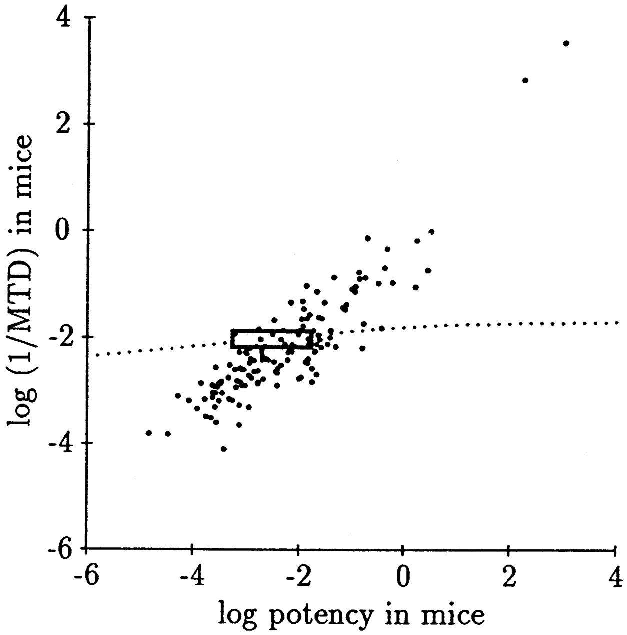

Figure 1: Assumptions in Piegorsch et al. (1992) Compared to NCI/NTP Data; Chemicals that are Statistically Significant Carcinogens in the Mouse; Logs to Base 10

4. New Simulations

This section presents results from new simulations, with more plausible assumptions. 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. The values x_m and x_r are the 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, $y_r = -\infty$.

Each "chemical" is subjected to the simulated NCI/NTP bioassay described in the previous section. The probability of cancer follows the one-hit model, equation (1), with a background cancer rate of $p_0 = 10\%$ and an upper bound of $p_{\max} = 90\%$. (Compare Shlyakhter et al., 1992, p. 78.) 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 (Section 2).

As before, chemicals are classified by the Cochran-Armitage Trend Test. After testing, a chemical is characterized by a set of "observed" values ($\hat{c}_m$, $\hat{c}_r$, x_m, x_r, $\hat{y}_m$, $\hat{y}_r$). 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, and $\hat{c}_m = 0$ otherwise, and similarly for $\hat{c}_r$; recall that x_m and x_r are log MTD's. Finally, $\hat{y}_m$ and $\hat{y}_r$ are the maximum likelihood estimates for log potency.

Each "chemical" (c_m, c_r, x_m, x_r, y_m, y_r) is generated as an independent and identically distributed observation from random variables C_m, C_r, X_m, X_r, Y_m, Y_r, ϵ_m, and ϵ_r. The variables C_m and C_r are carcinogenicity indicators. Conditioned on C_m and C_r, the log MTD variables X_m and X_r have a bivariate normal distribution with $corr(X_m, X_r) = .93$. (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). 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 $\epsilon = -\infty$. See (Freedman et al., 1993; Lin, 1994). In the simulations, we use the 0.005 level, one-sided; this closely matches classification by "authors' opinion" (Haseman, 1983b; Gold et al., 1989). In the NCI/NTP data, there were 53 chemicals significant at the .005 level in both species; Freedman et al. (1993) used the .025 level and found 87 chemicals significant in both species. (Changing levels from .005 to .025 in our simulations 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.)

Model A

We chose the parameters for Model A (Table 2) so that summary characteristics of simulated data would match the real NCI/NTP data, while observed concordance would overestimate true concordance: the bias is about 25 percentage points. The first row in Table 2 gives parameters for simulated chemicals that are "true" carcinogens in the mouse and in the rat ($C_m = C_r = 1$). As shown in the third column, this category has 20% of the probability. The remaining columns describe the conditional distribution for X_m, X_r, ϵ_m, and ϵ_r, given C_m and C_r. For example, given that $C_m = C_r = 1$, the log MTD for mice X_m is normally distributed with a mean of 2.0 and a standard deviation of 1.0; the log MTD for rats X_r is normally distributed with a mean of 1.6 and a standard deviation of 1.0; and so forth. The other three rows are read similarly; the dots in Table 2 indicate that the corresponding ϵ is $-\infty$. Recall that within each row, X_m and X_r have a correlation of .93, while ϵ_m and ϵ_r are independent of each other and of the pair (X_m, X_r). (Appendix A explains how parameters were chosen; logs are to base 10.)

In Model A, the variables C_m and C_r are independent, due to the choice of probabilities in Table 2. Specifically, the probability that a chemical is a rat carcinogen is 50%, whether or not it is a mouse carcinogen; likewise, the probability that a chemical is a mouse carcinogen is 40%, whether or not it is a rat carcinogen. Furthermore, for chemicals carcinogenic in both species, the yields ϵ_m and ϵ_r are independent. In that sense, mice and rats are qualitatively and quantitatively independent.

The primary statistic of interest is concordance. Classifying chemicals based on c_m and c_r gives a "true" 2 × 2 concordance table; classifying chemicals based on $\hat{c}_m$ and $\hat{c}_r$ gives an "observed" 2 × 2 concordance table. For each set of chemicals, the "true" and "observed" concordance tables are computed. In order to check on the realism of the simulation, we also compute the mean and standard deviation of the

Dist. of (C_m, C_r)			Dist. of X_m		Dist. of X_r		Dist. of ϵ_m		Dist. of ϵ_r	
c_m	c_r	prob.	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	1	.20	2.0	1.0	1.6	1.0	0.35	0.50	0.25	0.50
1	0	.20	2.3	0.9	1.8	0.9	−0.20	0.50	·	·
0	1	.30	2.1	1.5	1.8	1.5	·	·	−1.01	0.50
0	0	.30	2.7	1.0	2.2	0.9	·	·	·	·

Table 2: Parameters in Model A

log MTD variables x_m and x_r, for each of the four cells in the observed concordance table. For each of the two cells with $\hat{c}_m = 1$, we compute the mean and standard deviation of the estimated log potency $\hat{y}_m$. Likewise, for each of the two cells with $\hat{c}_r = 1$, we compute the mean and standard deviation of $\hat{y}_r$. Finally, for chemicals with $\hat{c}_m = 1$ and $\hat{c}_r = 1$, the correlations among x_m, x_r, $\hat{y}_m$, and $\hat{y}_r$ are computed.

Each simulated dataset contains 297 chemicals, the number of NCI/NTP bioassays. The whole procedure of generating, testing, and classifying a set of 297 chemicals is repeated 1000 times. At the end of each simulation, there are 1000 "true" and 1000 "observed" concordance tables; there are also 1000 sets of means and standard deviations; and 1000 correlation matrices. The results are averaged and compared to NCI/NTP data.

Results for Model A

Results are presented in Table 3. The left hand 2 × 2 table is the average of the 1000 "true" concordance tables in the simulation. For each set of 297 chemicals, the number of "true ++" chemicals is random; on average, 59.3 chemicals were truly "++", and the average true concordance was 50%. The right hand 2 × 2 table is the average of the observed concordance tables: on average, 52.8 chemicals were classified as "++", and the average observed concordance was 76%. The average observed concordance table from Model A was virtually identical to the observed concordance table for the NCI/NTP data (Table 1). The bias in observed concordance is about 25 percentage points, because the true concordance is 50%.

The MTD's and potencies generated according to Model A are very similar to NCI/NTP data. For example, consider the chemicals with statistically significant results in both species (observed "++"). Over 1000 sets of 297 simulated chemicals, the mean log MTD in mice of the observed ++'s averaged 2.00, and the standard deviation of the log MTD's averaged 1.00. In NCI/NTP data, the "++" chemicals have a mean log MTD in mice of 1.99 and a standard deviation of 1.02. See Table 4. Finally, the correlations among the simulated "++" chemicals closely match the correlations from NCI/NTP (Table 5). For results on the "+−", "−+", and "−−" chemicals, see Appendix B.

Model A: "True"				Model A: Observed			
		Rats				Rats	
		+	−			+	−
Mice	+	59.3	59.4	Mice	+	52.8	48.4
	−	89.4	89.0		−	22.1	173.8

Table 3: Concordance for 297 Chemicals tested both in Mice and Rats

	Model A		NCI/NTP	
	Average of Means	Average of SD's	Mean	SD
log MTD in mice	2.00	1.00	1.99	1.02
log potency in mice	−1.82	1.04	−1.80	1.09
log MTD in rats	1.60	1.00	1.60	1.02
log potency in rats	−1.47	1.04	−1.46	1.16

Table 4: Means and SD's for "++" Chemicals ($\hat{c}_m = \hat{c}_r = 1$)

Model A				
	X_m	X_r	$\hat{Y}_m$	$\hat{Y}_r$
X_m	1.00			
X_r	.93	1.00		
$\hat{Y}_m$	-.96	-.89	1.00	
$\hat{Y}_r$	-.89	-.96	.85	1.00

NCI/NTP				
	X_m	X_r	$\hat{Y}_m$	$\hat{Y}_r$
X_m	1.00			
X_r	.93	1.00		
$\hat{Y}_m$	-.92	-.85	1.00	
$\hat{Y}_r$	-.85	-.88	.86	1.00

Table 5: Correlations for "++" Chemicals ($\hat{c}_m = \hat{c}_r = 1$)

The Source of Bias in Concordance

It is natural to think that errors in classifying chemicals will cause concordance to go down, but this is not necessarily so. 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 Model A are presented in Table 6. The row totals give the average "true" number of each type of chemical, as reported at the left in Table 3. The column totals give the average "observed" number of each type of chemical, as reported at the right in Table 3.

On the average, over the 1000 sets of 297 chemicals, 59.3 were "true ++". Most of these (52.5) were observed as "++" in the simulated bioassays, but an average of 3.9+2.7=6.6 were misclassified as discordant ("+−" or "−+"). Also, 89.0 chemicals were "true −−"; of these, an average of .4+.4=.8 were misclassified as discordant. The average total number of "false discordances" can thus be computed from the first and fourth lines of the table as $3.9 + 2.7 + .4 + .4 = 7.4$. On the other hand, the average total number of "false concordances" is, from the second and third lines, $.2 + 15.4 + .1 + 70.0 = 85.7$. The number of false concordances is much larger than the number of false discordances: in particular, the "observed −−" cell is inflated,

True	**Observed**				
	++	+−	−+	−−	Total
++	52.5	3.9	2.7	.2	59.3
+−	.2	43.7	.1	15.4	59.4
−+	.1	.3	18.9	70.0	89.4
−−	.0	.4	.4	88.1	89.0
Total	52.8	48.4	22.1	173.8	297.0

Table 6: Simulation Results for Model A: Matrix of Classifications

due to lack of power in the bioassay. This is what makes the observed concordance much larger than the true concordance.

Model B

We chose the parameters for Model B (Table 7) so that summary characteristics of simulated data would match the real NCI/NTP data, while observed concordance would greatly overestimate true concordance. In Model B, all chemicals are carcinogenic in at least one species, but only 18% are carcinogenic in both species. Averaged over 1000 sets of 297 chemicals, the true concordance was 18%, and the observed concordance was 77%. As with Model A, the average observed concordance table was virtually identical to the concordance table for NCI/NTP. The MTD's, estimated potencies, and correlations generated according to Model B were similar to those for NCI/NTP data (Appendix B). In particular, Model B more or less fits the NCI/NTP data; yet mouse carcinogens are, in this model, much *less likely* than mouse noncarcinogens to be rat carcinogens—25% versus 100%.

Dist. of (C_m, C_r)			Dist. of X_m		Dist. of X_r		Dist. of ϵ_m		Dist. of ϵ_r	
c_m	c_r	prob.	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	1	.18	2.0	1.0	1.6	1.0	0.90	0.50	0.90	0.50
1	0	.53	2.3	0.9	1.8	0.9	−0.85	0.50	·	·
0	1	.29	2.1	1.5	1.8	1.5	·	·	−0.95	0.50

Table 7: Parameters in Model B

Model C

Our next simulation (Model C, Table 8) is designed to show that current bioassay design allows observed concordance in excess of 90%, with a true concordance even higher—by a little. In this simulation, means and SD's of log MTD and log potency match the real data reasonably well, as do the correlations (Appendix B); of course, the simulated observed concordance is much larger than the concordance seen in NCI/NTP data. As it turns out, the simulated observed concordance of 92% overestimates the "true" concordance in Model C, by about two percentage points.

Dist. of (C_m, C_r)			Dist. of X_m		Dist. of X_r		Dist. of ϵ_m		Dist. of ϵ_r	
c_m	c_r	prob.	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	1	.20	2.0	1.0	1.6	1.0	0.35	0.50	0.25	0.50
1	0	.05	2.3	0.9	1.8	0.9	−0.20	0.50	·	·
0	1	.05	2.1	1.5	1.8	1.5	·	·	−1.01	0.50
0	0	.70	2.4	1.0	2.0	0.9	·	·	·	·

Table 8: Parameters in Model C

Model D

Model D is characterized in Table 9. All chemicals are either carcinogenic in both species, or carcinogenic in neither species. The true concordance in Model D is 100%. Averaged over 1000 sets of 297 chemicals, the observed concordance was 77%; the average concordance table from Model D was virtually identical to the concordance table from NCI/NTP data. Furthermore, the MTD's, estimated potencies, and correlations generated according to Model D were similar to NCI/NTP data (Appendix B). Thus, the bias in observed concordance can be downward by a substantial amount, as suggested by Piegorsch et al. (1992).

Dist. of (C_m, C_r)			Dist. of X_m		Dist. of X_r		Dist. of ϵ_m		Dist. of ϵ_r	
c_m	c_r	prob.	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	1	.47	2.0	1.0	1.6	1.0	−0.24	0.50	−0.51	0.50
0	0	.53	2.5	1.0	2.1	1.0	·	·	·	·

Table 9: Parameters in Model D

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 Models A and B demonstrate, 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 (Model C). 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 in Model A, by computing concordance separately for chemicals with mouse MTDs above and below 100. (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 radically 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. Our simulations have several blemishes. For one thing, MTD's are even longer-tailed than the log normal. For another, we do not get enough variance in the $--$ cell (Tables 13, 15, 18, and 21 below). Furthermore, estimated NCI/NTP tumor yields show some dependence (Table 12 below), while true tumor yields are independent in the present model and estimated tumor yields are nearly uncorrelated. Correlation of yields can be built into the model, however.

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 Petō et al. (1984), *Cancer Research* (1991) Vol. 51 No. 23 Part 2 pp.6407–6491, 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, 1991, 1993, 1994), who discusses alternative models. Because of uncertainties about dose-response models, simulation 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.

Other literature

There have been many studies of concordance, either to validate species extrapolation or to analyze possible modifications of bioassay design. Some papers have been cited above. Also see, for instance, Griesemer and Cueto (1980), Purchase (1980), Haseman and Huff (1987), Haseman et al. (1987), Byrd, Crouch and Wilson (1990), Krewski, Goddard, and Withey (1990), Gold and Slone (1993), or Haseman and Lockhart (1993). Reproducibility of bioassay results is considered by Gold et al. (1987). For studies with a policy analysis flavor, see Lave et al. (1988), who use concordance data to argue that the current regulatory framework is not cost-effective; Ennever et al. (1990) consider the costs of uncertainties about concordance.

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 in 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 (1995).

5. Other Measures

There are many possible alternative measures to concordance. "Correlation" is the Pearson product moment correlation of the carcinogenicity indicators c_m and c_r. The true correlation is denoted $corr(c_m, c_r)$ and is estimated by $corr(\hat{c}_m, \hat{c}_r)$. For NCI/NTP data, the correlation is 0.45; see Table 10. Over 1000 sets of chemicals, the observed correlations in our four models averaged about 0.45, 0.45, 0.78, and 0.45, respectively. The "true" correlations averaged 0.0, −0.68, 0.73, and 1.00. (Models C and D were constructed so the true association would be strong.)

The "odds ratio" is defined as follows. Let n_{11} be the number of chemicals with $c_m = 1$ and $c_r = 1$; let n_{10} be the number of chemicals with $c_m = 1$ and $c_r = 0$; and so forth. Then

$$\text{true odds ratio} = \frac{n_{11}/n_{10}}{n_{01}/n_{00}} = \frac{n_{11}\ n_{00}}{n_{10}\ n_{01}}.$$

	Correlation		**Odds Ratio**	
	"True"	Observed	"True"	Observed
NCI/NTP	?	0.45	?	8.7
Model A	0.00	0.45	1.0	9.2
Model B	−0.68	0.45	0.0	9.5
Model C	0.73	0.78	56	107
Model D	1.00	0.45	∞	9.6

Table 10: Bias in Correlation and Odds Ratio

The corresponding estimator is

$$\text{observed odds ratio} = \frac{\hat{n}_{11}/\hat{n}_{10}}{\hat{n}_{01}/\hat{n}_{00}} = \frac{\hat{n}_{11}\ \hat{n}_{00}}{\hat{n}_{10}\ \hat{n}_{01}},$$

where $\hat{n}_{11}$ is the number of chemicals with $\hat{c}_m = 1$ and $\hat{c}_r = 1$, and so on. The odds ratio for NCI/NTP data is 8.7. The models gave average observed odds ratios of 9.2, 9.5, 107, and 9.6; however, the "true" odds ratios averaged 1.0, 0.0, 56, and $+\infty$. Correlation coefficients and odds ratios, like concordance, can be seriously biased; and the bias can go in either direction.

6. Inter-species Correlations of Carcinogenic Potency

Inter-species agreement can also be measured quantitatively. Crouch and Wilson (1979) observed a high inter-species correlation of log potencies among chemicals with statistically significant results in both mice and rats. Bernstein et al. (1985) demonstrated that this high correlation could be explained as a statistical artifact; also see (Freedman et al., 1993). This section reviews the arguments; the context is the NCI/NTP data discussed above.

The correlation of log potencies is 0.86 for the 53 NCI/NTP chemicals with statistically significant ($p \leq 0.005$, one-sided) potencies in both species; see the bottom right panel of Figure 2. To demonstrate the artifact in this correlation, suppose 10% of the animals in the control group of a standard NCI/NTP bioassay develop cancer. If the bioassay results are statistically significant and not all the dosed animals develop cancer, then the maximum likelihood estimate of log potency will be within 0.9 of log(1/MTD); that is,

(5) $\log(1/\text{MTD}) - 0.9 < \log(\text{estimated potency}) < \log(1/\text{MTD}) + 0.9.$

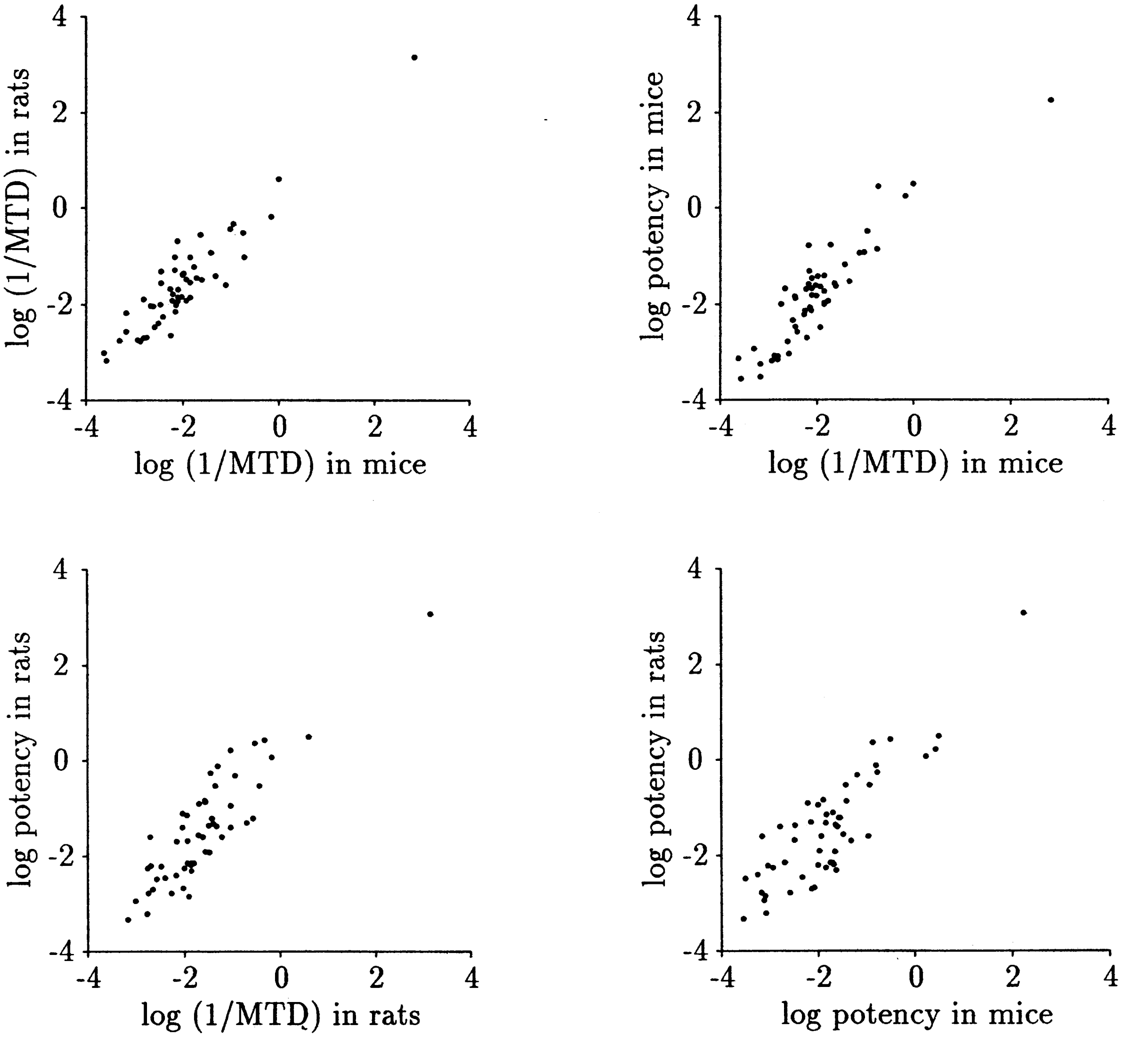

Figure 2: MTD's and Potencies; Chemicals with Statistically Significant Potencies in NCI/NTP Bioassays; Logs to Base 10

Inequality (5) is essentially the one in Bernstein et al. (1985), except the bounds are appropriate for NCI/NTP data (Lin, 1994); also see Gaylor and Gold (1995). In (5) and in Figure 2, logs are to base 10. Variability in control tumor rates may widen the bounds a little; so do lifetable adjustments to the estimates.

Let d be the MTD for a chemical, and let $\hat{b}$ be the estimated potency. Use the subscripts m and r to denote species. As shown in the top left panel of Figure 2, $\log(d_m) \approx \log(d_r)$. Equation (2) says that $-0.9 < \log(\hat{b}_m) + \log(d_m) < 0.9$ and $-0.9 < \log(\hat{b}_r) + \log(d_r) < 0.9$. Thus, $\log(\hat{b}_m) \approx -\log(d_m)$, as shown in the top right panel; also, $\log(\hat{b}_r) \approx -\log(d_r)$, as shown in the bottom left panel. It follows that $\log(\hat{b}_m) \approx \log(\hat{b}_r)$. This is the artifact, which explains the high correlation in the bottom right panel.

Of course, there is likely to be some non-artifactual relationship between carcinogenicity in mice and rats. Goodman and Wilson (1991, 1992) consider how the relationship between potencies depends on mutagenicity. Also, Freedman et al. (1993) found a weak relationship between tumor yields in the two species ($r \approx .45, n = 53$, Table 12 below).

If all the animals in the nonzero dose groups develop tumors at the same site, then the estimated potency is infinite. However, very few chemicals cause 100% tumor incidence. This might have some biological significance, but it might also point to other artifacts, such as errors in necropsy reports; for dicussion, see (Bernstein et al., 1985; Freedman et al., 1993; Krewski et al., 1993).

Crouch et al. (1987) responded to Bernstein et al. (1985) by claiming that the relationship between potency and MTD is based on biology, not statistics. Likewise, Goodman and Wilson (1992) observed that few chemicals have log(estimated potency $\times$ MTD)>1, and argued for biological significance; however, on the whole, this seems to be another manifestation of the artifact—and the absence of 100% tumor rates.

The MTD is barely sub-toxic. Therefore, animals in the high dose group may experience chronic cell killing and cell replacement, which tends to increase the risk of cancer. For this reason among others, toxicity may well be related to carcinogenicity. On the other hand, inequality (5) is a mathematical fact derived from bioassay design. It is the design of the experiments that precludes estimated potencies in the range (0, MTD/10) or (10 MTD, ∞). Moreover, the relationship between toxicity and carcinogenicity is a major complication in dose extrapolation, if tissue damage is much less—or much more easily repaired—at low environmental doses. For discussion of such issues, see Bernstein et al. (1985), Gold et al. (1989), Gold (1990), Ames and Gold (1990), Cunningham and Matthews (1991), Gold et al. (1992), Cohen and Ellwein (1992), Freedman et al. (1993), Parsons et al. (1995), or Ames, Gold, and Willett (1995); the last is a compact introduction to cancer biology and epidemiology. Bernstein et al. (1985) and Freedman et al. (1993) respond to the biological arguments made by Crouch et al. (1987) or Goodman and Wilson (1991).

Recent papers on quantitative inter-species agreement and the artifact include Whipple (1985), Rieth and Starr (1989), (Krewski et al., 1990, 1993); Kodell et al. (1991) discuss the role of the model. The extrapolation from rodents to humans is discussed in (Freedman and Zeisel, 1988), (Gold et al., 1989), (Gold et al., 1992); also see (Gaylor et al., 1993); Allen et al. (1988) take a more optimistic view, as do Goodman and Wilson (1991, 1992). Kodell et al. (1995) suggest that the observed interspecies correlation in potency may be biased low; they use the one-hit model, assuming further that (1) all chemicals are carcinogenic in both species, and (2) bioassays give unbiased estimates of potencies; in effect, measurement error attenuates the correlation. Our simulations indicate, however, that the situation may be more complicated than pure measurement error.

To avoid the artifact, various authors have suggested expanding the test set of chemicals, for instance, to include positive but insignificant potencies, or to replace estimates that are 0 by upper 95%-confidence limits, or to truncate estimates from below at small, positive values. None of these strategies seem to be effective: see Bernstein et al. (1985), Freedman et al. (1993), or Lin (1994).

7. Conclusions

For the NCI/NTP data, the observed concordance between mice and rats is about 75%. Simulation studies do not determine the direction of the bias in this estimate, but suggest that it can be substantial, and go in either direction. Thus, true concordance may be much higher than 75%, or much lower. Furthermore, previously reported quantitative correlations of interspecies potencies can be explained in terms of statistical artifact. In our present state of knowledge, it seems unlikely that true concordance can be determined from bioassay data.

Acknowledgments

We thank P. Bickel, L. Breiman, D. Gaylor, L. LeCam, W. Piegorsch, T. Slone, T. Speed and P. Switzer for helpful suggestions. Research of LSG partially 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.

References

Allen, B. C., Crump, K. S., and Shipp, A. M. (1988). Correlation between carcinogenic potency of chemicals in animals and humans. *Risk Analysis* **8** 531–544.

Ames, B. and Gold, L. (1990). Too many rodent carcinogens: mitogenesis increases mutagenesis. *Science* **249** 970–971. Correspondence in *Science* **250** 1644–1646 and **251** 10–13, 387–388, 606–608, and **252** 902–904.

Ames, B. N., Gold, L. S., and Willett, W. C. (1995). The causes and prevention of cancer. *Proceedings of the National Academy of Science, USA* **92** 5258–65

Bernstein, L., Gold, L., Ames, B., Pike M., and Hoel, D. (1985). Some tautologous aspects of the comparison of carcinogenic potency in rats and mice. *Fundamental and Applied Toxicology* **5** 79–86.

Bridges, B. A. (1981). *Carcinogenesis* **2** 371

Bridges, B. A. (1990). *Japanese Journal of Cancer Research* **81** 105

Byrd, D. M., Crouch, E. A., Wilson R. (1990). Do mouse liver tumors predict rat tumors? A study of concordance between tumors induced at different sites in rats and mice. *Progress in Clinical and Biological Research* **331** 19–41.

Cohen, S. M., Ellwein, L. B. (1992). Risk assessment based on high-dose animal exposure experiments. *Chemical Research in Toxicology* **5** 742–8.

Crouch, E. and Wilson, R. (1979). Inter-species comparison of carcinogenic potency. *Journal of Toxicology and Environmental Health* **5** 1095–1118.

Crouch, E., Wilson, R., and Zeise, L. (1987). Tautology or not tautology? *Journal* *of Toxicology and Environmental Health* **20** 1–10.

Cunningham, M. L. and Matthews, H. B. (1991). Relationship of hepatocarcinogenicity and hepatocellular proliferation induced by mutagenic noncarcinogens vs carcinogens. *Toxicology and Applied Pharmacology* **110** 505–113.

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.

Ennever, F. K., Rosenkranz H. S., Lave, L. B., Omenn, G. S. (1990). Value-of-information analysis of testing strategies: estimating the effect of uncertainty about the proportion of chemicals that are true human carcinogens. *Progress in Clinical and Biological Research* **340D** 295–304.

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

Freedman, D., Gold, L., and Slone, T. (1993). How tautologous are inter-species 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.

Gaylor, D. and Gold, L. (1995). Quick estimate of the regulatory virtually safe dose based on the maximum tolerated dose for rodent bioassays. Technical report, National Center for Toxicological Research, Jefferson, Arkansas.

Gold, L. (1990). The importance of data on mechanism of carcinogesis in efforts to predict low-dose human risk. *Risk Analysis* **13** 399–410.

Gold, L., 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., Manley, N., Slone, T., Garfinkel, G., Rohrbach, L., and Ames, B. (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., Manley, N., and Ames, B. (1992). Extrapolation of carcinogenicity between species: qualitative and quantitative factors. *Risk Analysis* **12** 579–588.

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

Gold, L. S. and Slone, T. H. (1993). Prediction of carcinogenicity from two versus four sex-species groups in the Carcinogenic Potency Database. *Journal of Toxicology and Environmental Health* **39** 143–157.

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

Gold, L. S., Wright, C. Bernstein, L. and de Veciana, M. (1987). Reproducibility of results in "near-replicate" carcinogenic bioassays. *Journal of the National Cancer Institute* **78** 1149-1158.

Goodman, G. and Wilson, R. (1991). Predicting the carcinogenicity of chemicals in humans from rodent bioassay data. *Environmental Health Perspectives* **94** 195–218.

Goodman, G. and Wilson, R. (1992). Comparison of the dependence of the TD_{50} on maximum tolerated dose for mutagens and nonmutagens. *Risk Analysis* **12** 525–533.

Griesemer, R. A. and Cueto, C. (1980). Toward a classification scheme for degrees of experimental evidence for the carcinogenicity of chemicals for animals. In *Molecular and Cellular Aspects of Carcinogen Screening Tests* (R. Montesano, H. Bartsch and L. Tomatis, eds.) International Agency for Research on Cancer, Lyons, France. Scientific Publication No. 27.

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 Huff, J. E. (1987). Species correlation in long-term carcinogenicity studies. *Cancer Letters* **37** 125–132

Haseman, J. K., Huff, J. E., Zeiger, E. and McConnell, E. E. (1987). Comparative results of 327 chemical carcinogenicity studies. *Environmental Health Perspectives* **74** 229–235.

Haseman, J. K., and Johnson, F. M. (1995). Analysis of rodent bioassay data for anticarcinogenic effects. *Mutation Research*, in press.

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.

Kodell, R., Basu, A., and Gaylor, D. (1995). On interspecies correlation of carcinogenic potencies. Technical report. National Center for Toxicological Research, Jefferson, Arkansas.

Kodell, R. L., Gaylor, D. W., and Chen J. J. (1991). Carcinogenic potency correlations: real or artifactual? *Journal of Toxicology and Environmental Health* **32** 1–9.

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.

Krewski, D., Goddard, M.J., Withey, J.R. (1990). Carcinogenic potency and interspecies extrapolation. *Progress in Clinical and Biological Research* **340D** 323–34.

Lave, L. B., Ennever, F. K., Rosenkranz, H. S., and Omenn, G. S. (1988). Information value of the rodent bioassay. *Nature* **336** 631–633.

Lin, T. (1994). Statistical issues in rodent bioassays: concordance and correlation. Ph.D. Department of Statistics, U. C. Berkeley.

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

Moolgavkar, S. H. (1991). Carcinogenesis models: an overview. *Basic Life Sciences* **58** 387–99, with discussion.

Moolgavkar, S. H. (1993). Cell proliferation and carcinogenesis models: general principles with illustrations from the rodent liver system. *Environmental Health Perspectives* **101 Suppl 5** 91–4.

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

Parsons, R., Li, G. M., Longley, M., Modrich, P. Liu, B. Berk, T., Hamilton, S. R., Kinzler, K. W., Vogelstein, B. (1995) Mismatch repair deficiency in phenotypically normal human cells. *Science* **268** 738–40.

Peto, R., Gray, R., Brantom P., and Grasso, P. (1984). Nitrosamine carcinogenesis in 5120 rodents: chronic administration of sixteen different concentrations of NDEA, NDMA, NPYR and NPIP in the water of 4440 inbred rats, with parallel studies on NDEA alone of the effects of age of starting (3, 6, or 20 weeks) and of species (rats, mice or hamsters). In *N-Nitroso Compunds: Occurrence, Biological Effects and Revelance to Human Cancer.* (I. K. O'Neill, R.C. von Borstel, C. T. Miller, J. Long, and H. Barsch, eds.) International Agency for Research on Cancer, Lyons, France. Scientific Publication No. 57, pp. 627–55.

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.

Peto, R., Pike, M., Bernstein, L., Gold, L., and Ames, B. (1984). 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.

Peto, R., Roe, F., Lee, R., Levy L. and Clack, J. (1975). Cancer and ageing in mice and men. *British Journal of Cancer* **32** 411–426.

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.

Purchase, I. F. (1980). Interspecies comparisons of carcinogenicity. *British Journal of Cancer* **41** 454–468.

Rieth, J. P. and Starr, T. B. (1989). Experimental design constraints on carcinogenic potency estimates. *Journal of Toxicology and Environmental Health* **27** 287–96.

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.

Shlyakhter, A., Goodman, G., and Wilson, R. (1992). Monte Carlo simulation of rodent carcinogenicity bioassays. *Risk Analysis* **12** 73–82.

Snedecor, G. and Cochran, W. (1967). *Statistical methods.* Iowa State University

Press, Ames, IA.

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.

Whipple, C. (1985). Toxicity and carcinogenicity. *Risk Analysis* **5** 261–4.

Appendix A

This section explains how parameters were chosen. Use the letters C, X, Y, and ϵ for random variables. The variable C indicates "true" carcinogenicity: $C = 1$ for "true" carcinogens, and $C = 0$ otherwise. The variable X stands for log MTD. The variable Y stands for true log potency; if a chemical is not a carcinogen, $Y = -\infty$. Finally, the letter ϵ stands for true log yield. For the carcinogens, $\epsilon = X + Y$; for the noncarcinogens, $\epsilon = -\infty$.

Use "hats" to denote observed values from the bioassay. Among the random variables, $\hat{C}$ indicates whether the chemical was an "observed" carcinogen (i.e., had a statistically significant trend at the .005 level), $\hat{Y}$ is the maximum likelihood estimate of log potency, and $\hat{\epsilon}$ is the maximum likelihood estimate of log yield. Among NCI/NTP chemicals, U is log MTD, $\hat{V}$ is estimated log potency, $\hat{D}$ indicates whether $\hat{V}$ is statistically significant, and $\hat{\delta} = U + \hat{V}$. (It is assumed that log MTD can be measured without error.) The notation is laid out in Table 11.

Use the subscripts m and r to denote species. Each of the pairs (ϵ_m, ϵ_r) and (X_m, X_r) is assumed to have a bivariate normal distribution. The pair (ϵ_m, ϵ_r) is assumed to be independent of the pair (X_m, X_r), that is, tumor yields are independent of MTDs. These assumptions are at least approximately true for real data. For example, for the 53 NCI/NTP chemicals with $\hat{D}_m = 1$ and $\hat{D}_r = 1$, the pair $(\hat{\delta}_m, \hat{\delta}_r)$ is approximately uncorrelated with the pair (U_m, U_r); see Table 12. (Of course, in real data, the "true" tumor yields are unobservable.)

Picking the parameters involves choosing the yields, the true concordance, and the MTD's. The first step was choosing parameters for the ϵ's. Given $C_m = 1$ and $C_r = 1$, the conditional expected value for ϵ_m was chosen by judgment, and likewise for the conditional expected value for ϵ_m given $C_m = 1$ and $C_r = 0$. Also, given $C_m = 1$ and $C_r = 1$, the conditional expected value for ϵ_r was chosen by judgment, and

	Model		NCI/NTP	
	Mice	Rats	Mice	Rats
“true” carcinogenicity	C_m	C_r		
“true” log MTD	X_m	X_r	U_m	U_r
“true” log potency	Y_m	Y_r		
“true” log yield	ϵ_m	ϵ_r		
statistical significance	$\hat{C}_m$	$\hat{C}_r$	$\hat{D}_m$	$\hat{D}_r$
estimated log potency	$\hat{Y}_m$	$\hat{Y}_r$	$\hat{V}_m$	$\hat{V}_r$
estimated log yield	$\hat{\epsilon}_m$	$\hat{\epsilon}_r$	$\hat{\delta}_m$	$\hat{\delta}_r$

Table 11: Notation

	U_m	U_r	$\hat{\delta}_m$	$\hat{\delta}_r$
U_m	1.00			
U_r	.93	1.00		
$\hat{\delta}_m$	.04	.03	1.00	
$\hat{\delta}_r$	−.07	−.01	.45	1.00

Table 12: Correlations in NCI/NTP Data, the 53 “++” Chemicals

likewise for the conditional expected value for ϵ_r given $C_m = 0$ and $C_r = 1$. (Initially, the conditional expectations for the ϵ's were set equal to the observed values from NCI/NTP data; for example, the conditional expectation for ϵ_m given $C_m = 1$ and $C_r = 1$ was set equal to the average value of $\hat{\delta}_m$ for chemicals with $\hat{D}_m = 1$ and $\hat{D}_r = 1$. The initial values for the conditional expectations were then modified by judgment.) Then, for chemicals with $C_m = 1$, the conditional standard deviation of ϵ_m was set at 0.5, which was the value for $\mathrm{SD}(\hat{\delta}_m|\hat{D}_m = 1)$, rounded to one decimal place. Likewise, for chemicals with $C_r = 1$, the conditional standard deviation of ϵ_r was set at 0.5, which was the value for $\mathrm{SD}(\hat{\delta}_r|\hat{D}_r = 1)$, again rounded to one decimal place. Finally, $corr(\epsilon_m, \epsilon_r)$ was set to zero.

The next step was determining the probabilities for the "true" concordance table. There are four possible values for the pair (C_m, C_r). Given a particular set of values for C_m and C_r, there are four possible classifications $(\hat{C}_m,\hat{C}_r)$. This gives rise to a 4 × 4 transition matrix. Call this matrix M; the ij^{th} entry of M gives the probability that a chemical of type i will be observed to be type j, where a type 1 chemical is "++", a type 2 chemical is "+−", and so forth. The matrix M controls the rate at which chemicals are misclassified. The various probabilities in M were found by numerical integration; of course, these depend on the mean and SD of the ϵ's, which control the power of the trend test. (Table 6 is an empirical analog to M in Model A, rescaled from probabilities to numbers.)

Let p be the row vector of proportions of NCI/NTP chemicals that are observed "++", "+−", "−+"; and "−−". For example, $p(++) = 53/297 = .178$; see Table 3. Let π be the row vector of probabilities for the model chemicals. The column vector π' for Model A is shown in Column 3 of Table 2. The row vector π was chosen for Model A as follows: first, π_0 was set equal to pM^{-1}; then π_0 was rounded slightly to achieve independence. For Models B and D, some elements of π_0 were slightly negative; these were truncated at zero, then π_0 was scaled and rounded so that the sum of entries was equal to 1. For Model C, the vector π was chosen by judgment.

The final step was determining parameters for the log MTD's. The conditional distribution for the X's was chosen as follows. Given C_m and C_r, the conditional correlation $corr(X_m, X_r)$ was set at .93; see Table 12. Next, SD$\{X_m | C_m = a$ and $C_r = b\}$ was set equal to the standard deviation of U_m for those NCI/NTP chemicals with $\hat{D}_m = a$ and $\hat{D}_r = b$. The conditional standard deviation for X_r was chosen similarly. If $a > 0$ or $b > 0$, then E$\{X_m | C_m = a$ and $C_r = b\}$ was set equal to the mean of U_m for those NCI/NTP chemicals with $\hat{D}_m = a$ and $\hat{D}_r = b$; likewise for X_r. For chemicals with $C_m = 0$ and $C_r = 0$, the conditional mean of X_m was chosen so that the unconditional mean E(X_m) would match the overall average of U_m for NCI/NTP chemicals; likewise for X_r. Finally, all the conditional means and conditional standard deviations for the X's were rounded to one decimal place.

Other things being equal, the observed concordance depends on the parameters for the true log yields ϵ_m and ϵ_r. Among truly "++" chemicals, if the true yields are both either very high or very low, the observed concordance is maximized; if one true yield is high and the other is low (say, ϵ_m is high and ϵ_r is low), then observed concordance goes down. In the true "+−" and "−+" cells, high true yields in one species lead to low observed concordance, and low true yields lead to high observed concordance (classification as "−−").

Appendix B

This section presents results for the simulations. As before, X_m is log MTD in mice, X_r is log MTD in rats, $\hat{Y}_m$ is estimated log potency in mice, and $\hat{Y}_r$ is estimated log potency in rats. The "++" part of Table 13 appears as Text Table 4.

	"++" Chemicals				"+−" Chemicals			
	Model A		NCI/NTP		Model A		NCI/NTP	
	Avg. of Means	Avg. of SD's	Means	SD	Avg. of Means	Avg. of SD's	Means	SD
Mice:								
log dose	2.00	1.00	1.99	1.02	2.27	0.91	2.28	0.86
log potency	−1.82	1.04	−1.80	1.09	−2.35	0.98	−2.30	1.05
Rats:								
log dose	1.60	1.00	1.60	1.02	1.79	0.91	1.80	0.90
log potency	−1.47	1.04	−1.46	1.16	·	·	·	·

	"−+" Chemicals				"−−" Chemicals			
	Model A		NCI/NTP		Model A		NCI/NTP	
	Avg. of Means	Avg. of SD's	Means	SD	Avg. of Means	Avg. of SD's	Means	SD
Mice:								
log dose	2.11	1.42	2.10	1.45	2.42	1.25	2.41	0.95
log potency	·	·	·	·	·	·	·	·
Rats:								
log dose	1.80	1.42	1.75	1.49	2.00	1.19	2.01	0.89
log potency	−2.13	1.47	−2.15	1.58	·	·	·	·

Table 13: Means and SD's for Model A

Model B: "True"				Model B: Observed				NCI/NTP			
		Rats				Rats				Rats	
		+	−			+	−			+	−
Mice	+	53.3	157.8	Mice	+	53.2	47.9	Mice	+	53	48
	−	85.9	0.0		−	21.8	174.1		−	22	174

Table 14: Concordance for 297 Chemicals tested both in Mice and Rats

	"++" Chemicals				"+−" Chemicals			
	Model B		NCI/NTP		Model B		NCI/NTP	
	Avg. of Means	Avg. of SD's	Means	SD	Avg. of Means	Avg. of SD's	Means	SD
Mice:								
log dose	2.00	1.00	1.99	1.02	2.29	0.90	2.28	0.86
log potency	−1.61	1.02	−1.80	1.09	−2.63	0.94	−2.30	1.05
Rats:								
log dose	1.60	1.00	1.60	1.02	1.79	0.90	1.80	0.90
log potency	−1.21	1.02	−1.46	1.16	·	·	·	·

	"−+" Chemicals				"−−" Chemicals			
	Model B		NCI/NTP		Model B		NCI/NTP	
	Avg. of Means	Avg. of SD's	Means	SD	Avg. of Means	Avg. of SD's	Means	SD
Mice:								
log dose	2.09	1.47	2.10	1.45	2.23	1.16	2.41	0.95
log potency	·	·	·	·	·	·	·	·
Rats:								
log dose	1.79	1.47	1.75	1.49	1.80	1.16	2.01	0.89
log potency	−2.15	1.49	−2.15	1.58	·	·	·	·

Table 15: Means and SD's for Model B

	Model B					NCI/NTP			
	X_m	X_r	$\hat{Y}_m$	$\hat{Y}_r$		X_m	X_r	$\hat{Y}_m$	$\hat{Y}_r$
X_m	1.00				X_m	1.00			
X_r	.93	1.00			X_r	.93	1.00		
$\hat{Y}_m$	–.98	–.91	1.00		$\hat{Y}_m$	–.92	–.85	1.00	
$\hat{Y}_r$	–.91	–.98	.90	1.00	$\hat{Y}_r$	–.85	–.88	.86	1.00

Table 16: Average Correlations for "++" Chemicals ($\hat{c}_m = \hat{c}_r = 1$) from Model B

Model C: "True"				Model C: Observed				NCI/NTP			
		Rats				Rats				Rats	
		+	−			+	−			+	−
Mice	+	58.9	14.7	Mice	+	52.0	15.8	Mice	+	53	48
	−	14.7	208.6		−	6.9	222.3		−	22	174

Table 17: Simulation Results for Model C: Concordance

	"++" Chemicals				"+−" Chemicals			
	Model C		NCI/NTP		Model C		NCI/NTP	
	Avg. of Means	Avg. of SD's	Means	SD	Avg. of Means	Avg. of SD's	Means	SD
Mice:								
log dose	2.00	1.00	1.99	1.02	2.22	0.94	2.28	0.86
log potency	−1.81	1.04	−1.80	1.09	−2.29	1.03	−2.30	1.05
Rats:								
log dose	1.60	0.99	1.60	1.02	1.76	0.93	1.80	0.90
log potency	−1.46	1.04	−1.46	1.16	·	·	·	·

	"−+" Chemicals				"−−" Chemicals			
	Model C		NCI/NTP		Model C		NCI/NTP	
	Avg. of Means	Avg. of SD's	Means	SD	Avg. of Means	Avg. of SD's	Means	SD
Mice:								
log dose	2.11	1.16	2.10	1.45	2.38	1.03	2.41	0.95
log potency	·	·	·	·	·	·	·	·
Rats:								
log dose	1.76	1.16	1.75	1.49	1.98	0.94	2.01	0.89
log potency	−1.85	1.04	−2.15	1.58	·	·	·	·

Table 18: Means and SD's for Model C

Model C	X_m	X_r	$\hat{Y}_m$	$\hat{Y}_r$
X_m	1.00			
X_r	.93	1.00		
$\hat{Y}_m$	-.96	-.89	1.00	
$\hat{Y}_r$	-.89	-.96	.86	1.00

NCI/NTP	X_m	X_r	$\hat{Y}_m$	$\hat{Y}_r$
X_m	1.00			
X_r	.93	1.00		
$\hat{Y}_m$	-.92	-.85	1.00	
$\hat{Y}_r$	-.85	-.88	.86	1.00

Table 19: Average Correlations for "++" Chemicals ($\hat{c}_m = \hat{c}_r = 1$) from Model C

Model D: "True"				Model D: Observed				NCI/NTP			
		Rats				Rats				Rats	
		+	−			+	−			+	−
Mice	+	139.6	0.0	Mice	+	53.1	47.4	Mice	+	53	48
	−	0.0	157.4		−	22.0	174.5		−	22	174

Table 20: Simulation Results for Model D: Concordance

	"++" Chemicals				"+−" Chemicals			
	Model D		NCI/NTP		Model D		NCI/NTP	
	Avg. of Means	Avg. of SD's	Means	SD	Avg. of Means	Avg. of SD's	Means	SD
Mice:								
log dose	2.00	1.00	1.99	1.02	2.02	1.00	2.28	0.86
log potency	−2.10	1.04	−1.80	1.09	−2.13	1.05	−2.30	1.05
Rats:								
log dose	1.60	1.00	1.60	1.02	1.62	1.00	1.80	0.90
log potency	−1.81	1.03	−1.46	1.16	·	·	·	·

	"−+" Chemicals				"−−" Chemicals			
	Model D		NCI/NTP		Model D		NCI/NTP	
	Avg. of Means	Avg. of SD's	Means	SD	Avg. of Means	Avg. of SD's	Means	SD
Mice:								
log dose	2.02	0.99	2.10	1.45	2.44	1.01	2.41	0.95
log potency	·	·	·	·	·	·	·	·
Rats:								
log dose	1.61	0.99	1.75	1.49	2.04	1.01	2.01	0.89
log potency	−1.85	1.04	−2.15	1.58	·	·	·	·

Table 21: Means and SD's for Model D

Model D

	X_m	X_r	$\hat{Y}_m$	$\hat{Y}_r$
X_m	1.00			
X_r	.93	1.00		
$\hat{Y}_m$	–.96	–.89	1.00	
$\hat{Y}_r$	–.89	–.96	.86	1.00

NCI/NTP

	X_m	X_r	$\hat{Y}_m$	$\hat{Y}_r$
X_m	1.00			
X_r	.93	1.00		
$\hat{Y}_m$	–.92	–.85	1.00	
$\hat{Y}_r$	–.85	–.88	.86	1.00

Table 22: Average Correlations for "++" Chemicals ($\hat{c}_m = \hat{c}_r = 1$) from Model D