Risk Analysis, Vol. 00, No. 0, 2014 DOI: 10.1111/risa.12314

Concordance of Noncarcinogenic Endpoints in Rodent Chemical Bioassays

Bing Wang[1] and George Gray[2,*]

Prediction of noncancer toxicologic outcomes in rodent bioassays of 37 chemicals from the National Toxicology Program was evaluated. Using the nonneoplastic lesions noted by NTP pathologists, we evaluate both agreement in toxic lesions across experiments and the predictive value of the presence (or absence) of a lesion in one group for other groups. We compare lesions between mice and rats, male mice and male rats, and female mice and female rats in both short-term and long-term bioassays. We also examine whether lesions found in a specific organ in a short-term test are also found in the long-term test of the same chemical. We find agreement (concordance) across species for specific lesions, as evaluated by the Kappa statistic, ranging from 0.58 (for concordance of nasal lesions between female mice and rats in long-term studies) to −0.14 (lung lesions between mice and rats in long-term studies). Predictive values are limited by the relatively small numbers of observations of each type of lesion. Positive predictive values range from 100% to 0%. Comparing the lesions found in short-term tests to those found in long-term tests resulted in Kappa statistic values from 0.76 (spleen lesions in male rats) to −0.61 (lung lesions in female mice). Positive predictive values of short-term tests for long-term tests range from 70% to 0%. Overall, there is considerable uncertainty in predicting the site of toxic lesions in different species exposed to the same chemical and from short-term to long-term tests of the same chemical.

KEY WORDS: Hazard identification; rodent bioassay; site concordance

1. INTRODUCTION

Toxicity testing in rodents is conducted to identify potential adverse health effects in humans from exposure to chemicals. Tests identify pathologic changes from chemical exposure, including damage to organs and alterations in physiologic functions. Rodent tests are used with the view that mammals (i.e., mice, rats, and humans) with similar anatomies and physiologies should respond similarly to exposures and testing can be done under well-controlled conditions. In traditional tests, animals are exposed to levels of a chemical considerably higher than those usually encountered by humans in order to elicit toxicologic responses. Exposure durations can be acute, subacute, subchronic, or chronic, lasting anywhere from 24 hours to two years.

There is an implicit assumption that animals are reliable predictors of human response. Testing the assumption of similar responses between rodents and humans is generally not possible because of a lack of human data of similar pathologic resolution. Some effort has been made comparing the results of preclinical animal tests of pharmaceuticals with the

[1]Food Safety and Technology Department, University of Nebraska, Lincoln, NE, USA.
[2]Department of Environmental and Occupational Health and Center for Risk Science and Public Health, George Washington University Milken Institute School of Public Health, Washington, DC, USA.
*Address correspondence to George Gray, Department of Environmental and Occupational Health, Center for Risk Science and Public Health, GWU Milken Institute School of Public Health, 950 New Hampshire Ave., NW, Washington, DC 20052, USA; gmgray@gwu.edu.

adverse events identified in clinical trials[1] but similar efforts for general chemical exposures are limited. Studies have assessed concordance between mice and rats for neoplastic endpoints[2,3] but few if any studies have been done on noncarcinogenic effects. The concordance between mice and rats tested under similar circumstances probably represents an upper bound on the concordance expected between rodents and humans.

In this study we compare the nonneoplastic lesions identified by pathologists as statistically or biologically significant in rodent bioassays conducted through the National Toxicology Program of the U.S. National Institute of Environmental Health Sciences. These studies are of very high quality and conducted under a rigorous set of protocols, including pathology procedures and definitions, which should make them ideal for such comparisons.[4] We evaluate the concordance of pathologic responses between mice and rats exposed to the same chemical under the same experimental conditions in both subchronic bioassays and chronic bioassays. We also look at the concordance of response within a sex/species between subchronic and chronic tests. In addition, we look at the predictive value of the presence, or lack, of a pathologic lesion in one group to other rodents exposed to the same chemical.

2. METHODS

2.1. Study Chemicals and Data Collection

The National Toxicology Program, an interagency program created to coordinate, strengthen, develop, and improve science-based toxicology testing, and provide up-to-date toxicological information to the public, has developed a thorough database of technical reports on long- and short-term toxicity studies. The studies are performed under the direction of the National Institute of Environmental Health Sciences (NIEHS), and are conducted in compliance with NTP laboratory health and safety requirements, meet all applicable health and safety regulations, animal care and use regulations, and Food and Drug Administration (FDA) Good Laboratory Practice Guidelines.[4,5] Studies are subjected to retrospective quality assurance audits before final publication. The studies evaluate the toxicologic potential of selected chemicals in laboratory rodents and provide details on neoplastic (in two-year bioassays) and nonneoplastic effects in control and treated animals. Chemicals studied are chosen primarily on the bases of human exposure, level of production, and chemical structure. NTP does not extrapolate these results to other species or conduct quantitative risk analysis for humans.

For this preliminary analysis, the 39 NTP studies conducted between 2000 and 2013 with both short- and long-term studies conducted on male and female mice and rats by gavage, feed, or drinking water were utilized to enhance comparability. Short-term studies lasted around three months, ranging between 12 and 14 weeks in length, and long-term studies lasted two years. B6C3F1 and B6C3F1/N mouse strains and F344/N and Wistar rat strains were used in the bioassays. Two chemicals were dropped from the analysis because they did not report any nonneoplastic lesions in the short- or long-term bioassays, resulting in a total of 37 chemicals. From the NTP Technical Reports we collected chemical name, chemical abstract service registry number (CAS No.), route of exposure, length of study, species, sex, organ where nonneoplastic lesions were identified, and nonneoplastic lesion type at any dose. For each chemical/species/sex combination, the attributed lesions were those recorded having significant increases in dosed groups compared to controls for short-term exposure and those identified in the "Nonneoplastic Effects" section of the summary tables in the abstract of each Technical Report for long-term exposure. In this way only lesions (rounded up to the organ level) judged statistically or biologically significant by study pathologists were included. All chemicals are identified by CAS number. Goldenseal Root Powder was not assigned a CAS number, so we labeled it "goldenseal." For this analysis, B6C3F1 and B6C3F1/N mouse strains were combined as "mice" and the F344/N and Wistar rat strains combined as "rats." Due to differences in the labeling of NTP entries over time, Forestomach, Stomach/Forestomach, Glandular Stomach, and Stomach were grouped under the group "Stomach," Mesenteric and Mandibular Lymph Nodes were grouped as "Lymph Node," and Large Intestine, Rectum, and Cecum were grouped as "Large Intestine" based on the descriptions available in the NTP report labeling all microscopic terms by organ.[6]

2.2. Characterization of Nonneoplastic Concordance

Concordance of nonneoplastic pathologic responses was measured in two ways: (1) between mice

Table I. Parameters Describing Nonneoplastic Lesion Concordance

	Qualitative	Quantitative
Nondirectional	*p* value from independence tests; Kappa agreement category	Concordance percentage; Kappa value
Directional	N/A	Positive predictive value (PPV); Negative predictive value (NPV)

and rats (including by sex), and (2) between short- and long-term exposures. For each comparison, concordances were evaluated qualitatively and quantitatively both with and without "directionality" (Table I). To conduct the analyses we used the statistical computation program R (Version 3.0.0) to generate a series of 22 tables to compare nonneoplastic chemical effects across species and across exposure lengths, once the selected data were collected and organized for all 37 chemicals. As in Fig. 1, chemicals that produced a positive response for a specific tissue or organ in both mice and rats, in mice but not in rats, not in mice but in rats, or neither in mice nor in rats were placed in the cells of $Mice^{+}/Rats^{+}$, $Mice^{+}/Rats^{-}$, $Mice^{-}/Rats^{+}$, and $Mice^{-}/Rats^{-}$. The 22 tables for interexposure-length comparisons were generated in a similar way.

For concordance across species, we evaluated the ability of lesions identified in a specific organ in mice to predict lesions in the same organ in rats and *vice versa*, which we characterize as dual-directional prediction. Similarly for concordance across exposure length, the directional prediction was measured from short- to long-term exposure because the primary question of interest was the application of short-term studies to predict the long-term toxicologic responses. Positive predictive value (PPV) and negative predictive value (NPV) were also computed for quantifying directional prediction (Fig. 1(a)). Mouse-to-rat PPV is defined as the percentage of chemicals for which lesions observed in a specific organ in response to a specific chemical in mice are also observed in rats exposed to the same chemical. Mouse-to-rat NPV is defined as the percentage of chemicals for which lesions are not observed in a specific organ in mice and also not observed in the same organ in rats when exposed to the same chemical. Rat-to-mouse PPV and rat-to-mouse NPV were defined in the same way. Fig. 1(b) shows the calculation of short-to-long PPV and short-to-long NPV.

Interspecies and interexposure lengths concordances were nondirectionally quantified by concordance percentage and Kappa value, and evaluated by independence tests and Kappa agreement category.[7] Concordance percentage, take interspecies concordance, for example, was defined as the percentage of chemicals placed in the $Mice^{+}/Rats^{+}$ and $Mice^{-}/Rats^{-}$ cells out of the 37 chemicals. The Kappa value is widely used as a statistical measure of interrater agreement for qualitative (ordinal or binominal) items, which is generally thought to be a more accurate measure than a simple percent agreement (equivalent to concordance percentage in our case) due to the consideration of randomly occurring agreement.[8] We therefore also applied the Kappa value to quantify the interspecies and interexposure lengths agreement in pathologic lesion response. The following equations show how interspecies Kappa was computed and applied. The identical approach was applied to interexposure lengths Kappa calculations.

Kappa values were then interpreted by commonly used agreement categories: <0 for poor agreement, 0.01–0.20 for slight agreement, 0.21–0.40 for fair agreement, 0.41–0.60 for moderate agreement, 0.61–0.80 for substantial agreement, and 0.81–0.99 for almost perfect agreement.[7] Perfect agreement would equate to a Kappa of 1 and chance agreement would equate to 0. We used the Kappa agreement category to qualify nondirectional concordance as an ordinal measure. Nondirectional concordance was also qualitatively evaluated by testing the independence across species and across exposure lengths in organ-specific lesion occurrence. The independence between mice and rats or between short- and long-term exposures was tested by the Chi-square test or Fisher's exact test when the Chi-square test was inappropriate. The general rule for using the Chi-square test is that the smallest expected frequency among the four cells of 2×2 tables should be at least 5.[9] Statistical significance of independence was considered at a p value less than 0.05. For example, a p value less than 0.05 in an independence test across animal species indicated a statistically significant association, meaning that knowing the occurrence of endpoints in one animal

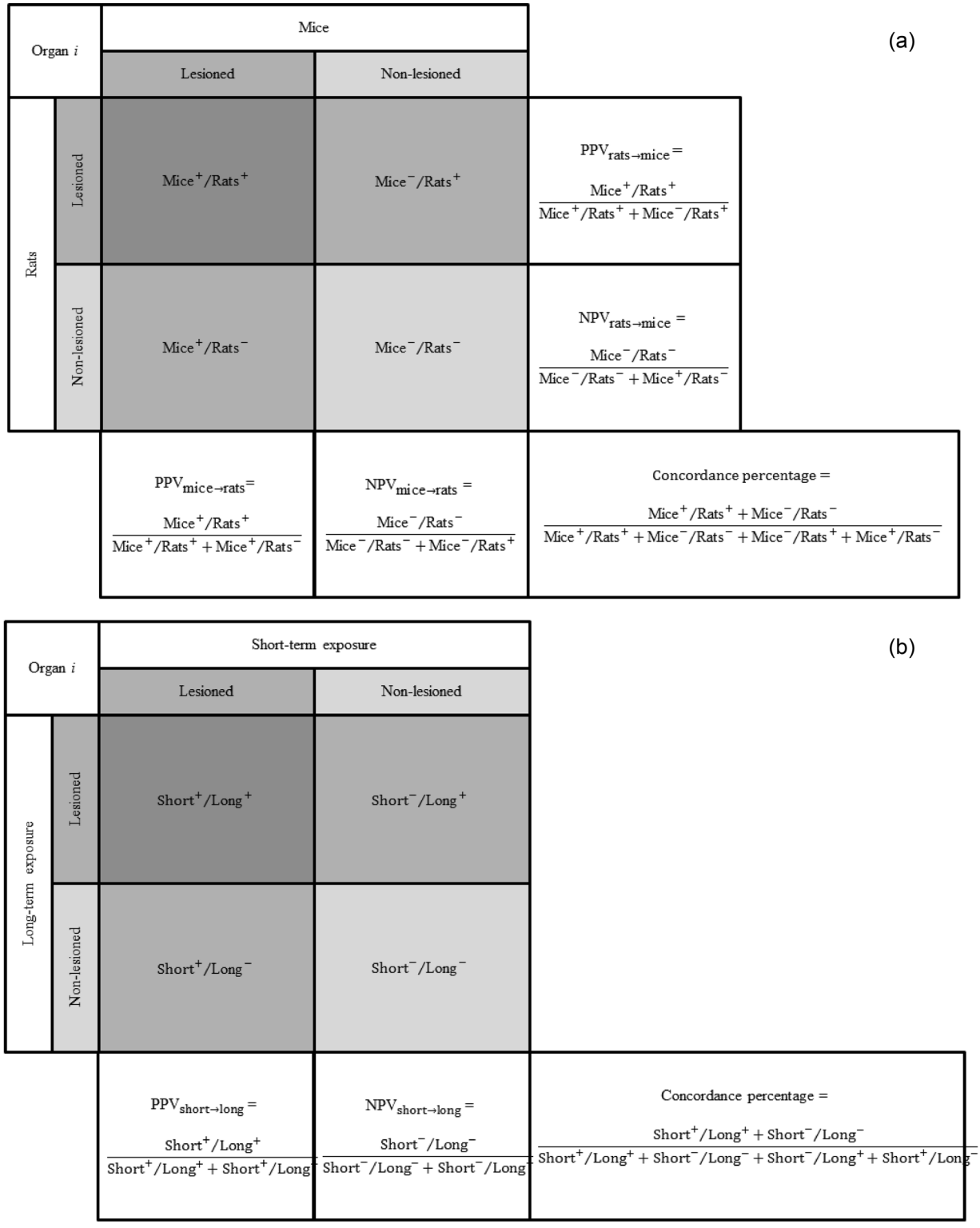

Fig. 1. Illustration of calculation of positive predictive value (PPV), negative predictive value (NPV), and concordance percentage. Organ *i* included bone marrow, kidney, liver, lung, nose, spleen, and stomach. (a) Demonstrates rat-to-mouse and mouse-to-rat predictions, (b) illustrates predictions from short-term to long-term tests of the same chemical.

Table II. Chemicals in This Analysis (37): From NTP Database with Short- and Long-Term Tests and Administered by Oral Routes Between 2000 and 2013

Chemical	CAS No.	Exposure Route	Report Year
2,4-Hexadienal	142–83–6	Gavage	2003
2-Methylimidazole	693–98–1	Feed	2004
4-Methylimidazole	822–36–6	Feed	2007
5-(Hydroxymethyl)-2-Furfural	67–47–0	Gavage	2010
Acrylamide	79–06–1	Drinking water	2012
Acrylonitrile	107–13–1	Gavage	2001
α,β-Thujone	76231–76–0	Gavage	2011
Androstenedione	63–05–8	Gavage	2010
Anthraquinone	84–65–1	Feed	2005
Benzophenone	119–61–9	Feed	2006
β-Myrcene	123–35–3	Gavage	2010
Bromochloroacetic Acid	5589–96–8	Drinking water	2009
Citral	5392–40–5	Feed	2003
Dibromoacetic Acid	631–64–1	Drinking water	2007
Dibromoacetonitrile	3252–43–5	Drinking water	2010
Dipropylene Glycol	25265–71–8	Drinking water	2004
Elmiron®	37319–17–8	Gavage	2004
Emodin	518–82–1	Feed	2001
Formamide	75–12–7	Gavage	2008
Ginkgo Biloba Extract	90045–36–6	Gavage	2013
Goldenseal Root Powder	GOLDENSEALRT	Feed	2010
Isoeugenol	97–54–1	Gavage	2010
Kava Kava Extract	9000–38–8	Gavage	2012
Methacrylonitrile	126–98–7	Gavage	2001
Methylene Blue Trihydrate	7220–79–3	Gavage	2008
Methyleugenol	93–15–2	Gavage	2000
Milk Thistle Extract	84604–20–6	Feed	2011
N,N-Dimethyl-P-Toluidine	99–97–8	Gavage	2012
o-Nitrotoluene	88–72–2	Feed	2002
p,p'-Dichlorodiphenyl Sulfone	80–07–9	Feed	2001
p-Nitrotoluene	99–99–0	Feed	2002
Primidone	125–33–7	Feed	2000
Pulegone	89–82–7	Gavage	2011
Pyridine	110–86–1	Drinking water	2000
Riddelliine	23246–96–0	Gavage	2003
Sodium Dichromate Dihydrate	7789–12–0	Drinking water	2008
Sodium Nitrite	7632–00–0	Drinking water	2001
trans-Cinnamaldehyde (microencapsulated)	14371–10–9	Feed	2004

species could help predict the occurrence in the other species.

3. RESULTS

As listed in Table II, a total of 37 chemicals were included in our analysis to examine the concordance in producing nonneoplastic lesions between rodent species (mice vs. rats) and between different exposure lengths of toxicologic studies (short-term and long-term) under identical experimental conditions. The list of 37 chemicals was finalized by excluding two chemicals (Chromium picolinate monohydrate [CAS No. 27882–76–4]; Ginseng [CAS No. 50647–08–0]) since no lesions were identified in either short-term or long-term studies. Our analysis revealed that chemicals produced nonneoplastic lesions at least once in 41 organs. Of the 41 organs, seven stood out with relatively high chemical response rates (greater than 15% of chemicals). The chemical response rate in a given organ was defined as the percentage of chemicals observed producing at least one pathologic lesion in the organ. To ensure sufficient numbers of responses for comparisons, we focused on the seven

organs with the highest number of responses: bone marrow, kidney, liver, lung, nose, spleen, and stomach.

As shown in Table III, liver and kidney had the highest chemical response rates, followed by spleen, stomach, nose, bone marrow, and lung. Nonneoplastic lesions were identified in liver in at least one sex/species and dose group for approximately three-quarters of studied chemicals (27/37, 72.97%), and in kidney for over half the chemicals (23/37, 62.16%) in either short-term or long-term studies. Liver and kidney had highest response rates in both short-term and long-term studies. Higher response rates for both liver and kidney were observed in the long-term compared to the short-term studies, though the difference did not demonstrate statistical significance. Spleen, stomach, nose, bone marrow, and lung lesions were found with smaller numbers of chemicals, ranging from 11 to 17. Rats appeared to show more lesions in liver, kidney, or bone marrow than mice. Response rates were comparable between female and male rodents in all listed organs.

Prediction of organ-specific lesions between rodent species and between exposure lengths were evaluated by PPV, NPV, concordance, Kappa value (and Kappa value agreement category[7]), and the *p* value from Chi-square or Fisher's exact test, which were all computed based on 22 corresponding tables. Cell values in the 22 tables and results of the concordance evaluation parameters are presented in Tables IV and V. In the following two sections, the lesion concordance between rodents exposed to the same chemical and between exposure lengths are illustrated, respectively.

Table III. Organs Demonstrating Highest Rates of Lesions and Used in This Analysis by Animal Species, Sex, and Chemical Exposure Length

Organ	Short-Term Studies								Long-Term Studies								Short- and Long-Term Studies							
	Mouse			Rat					Mouse				Rat				Mouse				Rat			
	F	*M*	*F&M*	*F*	*M*	*F&M*		Overall number (response rate)	*F*	*M*	*F&M*		*F*	*M*	*F&M*		*F*	*M*	*F&M*		*F*	*M*	*F&M*	Overall number (response rate)
Bone Marrow	1	2	2	8	7	8		9(24.32%)	1	1	2		4	3	5		6(16.22%)	2	3		4	9	8	10	12(32.43%)
Kidney	3	3	3	13	14	15		17(45.95%)	7	7	9		10	13	14		18(48.65%)	9	8		10	16	18	19	23(62.16%)
Liver	13	13	13	18	17	19		21(56.76%)	16	17	18		21	22	23		25(67.57%)	18	18		19	24	23	25	27(72.97%)
Lung	1	1	1	3	2	3		4(10.81%)	4	3	5		4	2	4		9(24.32%)	4	3		5	6	3	6	11(29.73%)
Nose	7	7	7	7	7	7		10(27.03%)	8	8	8		7	8	8		11(29.73%)	10	10		10	9	9	9	13(35.14%)
Spleen	7	7	7	8	8	9		14(37.84%)	8	8	10		8	7	9		13(35.14%)	11	11		13	10	9	11	17(45.95%)
Stomach	5	5	5	8	7	9		11(29.73%)	8	8	10		5	7	8		11(29.73%)	11	10		12	11	10	13	17(45.95%)

3.1. Concordance Across Species

We used concordance percentage and the Kappa value to quantify the interspecies nondirectional concordance, Kappa agreement category to qualify the concordance as an ordinal measure, and Chi-square or Fisher's exact test as a binominal measure. Compared to short-term studies, long-term studies showed a higher nondirectional concordance between mice and rats in producing organ-specific lesions. The interspecies concordance percentages ranged from 57% to 89% with an average of 75% in short-term studies and from 65% to 89% with an average of 80% in long-term studies. As shown in Table IV, the values for concordance percentage were heavily influenced by the large values of the negative/negative cells in the 22 tables. We therefore

Table IV. Noncancer Lesion Concordance Parameters: Cell Values, PPV, NPV, Concordance Percentage, *p* Value, Kappa Value, and Kappa Agreement Category

Organ	+/+	+/−	−/+	−/−	Mice to Rats PPV	Mice to Rats NPV	Rats to Mice PPV	Rats to Mice NPV	Concordance Percentage	*p* value	Kappa	Agreement Category
Mouse and Rat Concordance												
Short-term studies												
Bone marrow	1	1	7	28	50.00%	80.00%	12.50%	96.55%	78.38%	0.39*	0.12	Small
Kidney	1	2	14	20	33.33%	58.82%	6.67%	90.91%	56.76%	1.00*	−0.03	No
Liver	11	2	8	16	84.62%	66.67%	57.89%	88.89%	72.97%	0.01	0.46	Moderate
Lung	0	1	3	33	0.00%	91.67%	0.00%	97.06%	89.19%	1.00*	−0.04	No
Nose	4	3	3	27	57.14%	90.00%	57.14%	90.00%	83.78%	0.01*	0.47	Moderate
Spleen	2	5	7	23	28.57%	76.67%	22.22%	82.14%	67.57%	1.00*	0.05	Small
Stomach	3	2	6	26	60.00%	81.25%	33.33%	92.86%	78.38%	0.08*	0.31	Fair
Long-term studies												
Bone marrow	1	1	4	31	50.00%	88.57%	20.00%	96.88%	86.49%	0.26*	0.23	Fair
Kidney	5	4	9	19	55.56%	67.86%	35.71%	82.61%	64.86%	0.25*	0.20	Small
Liver	16	2	7	12	88.89%	63.16%	69.57%	85.71%	75.68%	0.00	0.52	Moderate
Lung	0	5	4	28	0.00%	87.50%	0.00%	84.85%	75.68%	1.00*	−0.14	No
Nose	5	3	3	26	62.50%	89.66%	62.50%	89.66%	83.78%	0.01*	0.52	Moderate
Spleen	6	4	3	24	60.00%	88.89%	66.67%	85.71%	81.08%	0.01*	0.50	Moderate
Stomach	7	3	1	26	70.00%	96.30%	87.50%	89.66%	89.19%	0.00*	0.71	Substantial
Female Mice and Female Rats												
Short-term studies												
Bone marrow	1	0	7	29	100.00%	80.56%	12.50%	100.00%	81.08%	0.22*	0.18	Small
Kidney	1	2	12	22	33.33%	64.71%	7.69%	91.67%	62.16%	1.00*	−0.01	No
Liver	11	2	7	17	84.62%	70.83%	61.11%	89.47%	75.68%	0.00	0.51	Moderate
Lung	0	1	3	33	0.00%	91.67%	0.00%	97.06%	89.19%	1.00*	−0.04	No
Nose	4	3	3	27	57.14%	90.00%	57.14%	90.00%	83.78%	0.01*	0.47	Moderate
Spleen	2	5	6	24	28.57%	80.00%	25.00%	82.76%	70.27%	0.63*	0.08	Small
Stomach	3	2	5	27	60.00%	84.38%	37.50%	93.10%	81.08%	0.06*	0.35	Fair
Long-term studies												
Bone marrow	1	0	3	33	100.00%	91.67%	25.00%	100.00%	91.89%	0.11*	0.37	Fair
Kidney	4	3	6	24	57.14%	80.00%	40.00%	88.89%	75.68%	0.07*	0.32	Fair
Liver	13	3	8	13	81.25%	61.90%	61.90%	81.25%	70.27%	0.02	0.42	Moderate
Lung	0	4	4	29	0.00%	87.88%	0.00%	87.88%	78.38%	1.00*	−0.12	No
Nose	5	3	2	27	62.50%	93.10%	71.43%	90.00%	86.49%	0.00*	0.58	Moderate
Spleen	5	3	3	26	62.50%	89.66%	62.50%	89.66%	83.78%	0.01*	0.52	Moderate
Stomach	3	5	2	27	37.50%	93.10%	60.00%	84.38%	81.08%	0.06*	0.35	Fair
Male Mice and Male Rats												
Short-term studies												
Bone marrow	1	1	6	29	50.00%	82.86%	14.29%	96.67%	81.08%	0.35*	0.15	Small
Kidney	1	2	13	21	33.33%	61.76%	7.14%	91.30%	59.46%	1.00*	−0.02	No
Liver	11	2	6	18	84.62%	75.00%	64.71%	90.00%	78.38%	0.00	0.56	Moderate
Nose	5	3	3	26	62.50%	89.66%	62.50%	89.66%	83.78%	0.01*	0.52	Moderate
Spleen	3	5	4	25	37.50%	86.21%	42.86%	83.33%	75.68%	0.16*	0.25	Fair
Stomach	4	4	3	26	50.00%	89.66%	57.14%	86.67%	81.08%	0.03*	0.42	Moderate

(*Continued*)

applied the Kappa value and Kappa agreement category to evaluate lesion concordance with a measure that reduces the influence of a single cell value. Among the seven organs, Kappa values ranged from −0.04 (poor agreement) to 0.47 (moderate agreement) with an average of 0.19 (slight agreement) in short-term studies and from −0.14 (poor agreement) to 0.71 (substantial agreement) with an average

Table IV. (Continued)

Organ	+/+	+/−	−/+	−/−	Mice to Rats PPV	Mice to Rats NPV	Rats to Mice PPV	Rats to Mice NPV	Concordance Percentage	*p* value	Kappa	Agreement Category
Long-term studies												
Bone marrow	0	1	3	33	0.00%	91.67%	0.00%	97.06%	89.19%	1.00*	−0.04	No
Kidney	5	2	8	22	71.43%	73.33%	38.46%	91.67%	72.97%	0.07*	0.34	Fair
Liver	15	2	7	13	88.24%	65.00%	68.18%	86.67%	75.68%	0.00	0.52	Moderate
Lung	0	3	2	32	0.00%	94.12%	0.00%	91.43%	86.49%	1.00*	−0.07	No
Nose	5	3	3	26	62.50%	89.66%	62.50%	89.66%	83.78%	0.01*	0.52	Moderate
Spleen	3	5	4	25	37.50%	86.21%	42.86%	83.33%	75.68%	0.16*	0.25	Fair
Stomach	4	4	3	26	50.00%	89.66%	57.14%	86.67%	81.08%	0.03*	0.42	Moderate

Notes: The *p* values were computed from either Fisher's exact tests or Chi-squared tests for independence. Fisher's exact was used if any cell contained fewer than five chemicals. Entries with (*) indicate *p* values were computed from Fisher's exact test; otherwise, the Chi-square test was used.

of 0.36 (fair agreement) in long-term studies. In short-term studies, the nose was the organ with the highest Kappa value, followed by liver, stomach, bone marrow, spleen, kidney, and lung. In long-term studies, the order was stomach, nose, liver, spleen, bone marrow, kidney, and lung. Chi-square or Fisher's exact tests also showed similar results; more organs with small *p* values (less than 0.05) in long-term than short-term studies.

In Table IV, a total of 42 concordance percentages and agreement categories for organ-specific endpoints were available. Based on the 42 paired data, Fig. 2 was generated to explore the correlation between two approaches for measuring the concordance of endpoints across species: (1) concordance percentages that are commonly used to evaluate cancerous endpoints across species and (2) agreement categories based on Kappa values that were introduced in this study to eliminate the strong influence of the negative/negative cell. As shown in Fig. 2, the concordance percentages fell into a wide range from 56.76% to 91.89% for "No" agreement category. The minimum concordance percentage for "Small" agreement was 64.86% and the distribution of concordance percentages for "Fair" agreement shifted more to the right compared to the "Small" agreement. The "Fair" and "Moderate" agreement categories had a similar spreading of concordance percentages, but the "Moderate" distribution had a larger median.

Interspecies directional concordance was quantified by PPV and NPV for producing organ-specific lesions due to chemical exposure. PPV and NPV were conducted for both mouse-to-rat and rat-to-mouse prediction. Mouse-to-rat PPVs were consistently higher across organs in long-term studies, ranging from 0% to 88.9%, with an average of 55.3%, compared to short-term studies ranging from 0% to 84.6% with an average of 44.8%. In both short-term and long-term studies, mouse-to-rat PPVs were highest in liver, nose, and stomach. Female mouse-to-rat and male mouse-to-rat PPVs are identical in value across organs within short-term and similar in value within long-term studies, except for bone marrow. Bone marrow short-term mouse-to-rat PPVs were 100% among females and 50% among males, while the difference was even larger in long-term studies with 100% among females and 0% among males. However, the large difference in PPVs was actually a result of the change in only one chemical producing the lesion. For example, as shown in Table III, only one chemical produced lesions in bone marrow in female mice and this very chemical also produced lesions in bone marrow in female rats due to short-term chemical exposure, which led to a mouse-to-rat PPV of 100%. Again in short-term studies, however, bone marrow lesions were observed in male mice for two chemicals and one of the two produced bone marrow pathology in male rats, which led to a mouse-to-rat PPV of 50%. A similar situation applies in the evaluation of long-term studies including the difference in female and male mouse-to-rat PPVs.

Rat-to-mouse PPVs were similar or lower than mouse-to-rat PPVs across organs. Rat-to-mouse PPVs are comparable to those of mouse-to-rat for liver, nose, and stomach, and 0% in lung both within females and within males. A clear decrease in rat-to-mouse PPVs was observed in kidney and bone marrow. In kidney, for example, in short-term studies, rat-to-mouse PPVs were 7.7% for females and 7.1% for males, whereas mouse-to-rat PPVs were 33.3% in

Table V. Short-Term to Long-Term Bioassay Lesion Concordance Parameters: Cell Values, PPV, NPV, Concordance Percentage, *p* Value, Kappa Value, and Agreement Category

Organ	+/+	+/−	−/+	−/−	Short-to-Long		Concordance Percentage	*p* Value	Kappa	Agreement Category
					PPV	NPV				
Mice										
Bone marrow	0	2	2	33	0.00%	94.29%	89.19%	1.00*	−0.06	No
Kidney	2	1	7	27	66.67%	79.41%	78.38%	0.14*	0.24	Fair
Liver	12	1	6	18	92.31%	75.00%	81.08%	0.00	0.62	Substantial
Lung	1	0	4	32	100.00%	88.89%	89.19%	0.14*	0.30	Fair
Nose	5	2	3	27	71.43%	90.00%	86.49%	0.00*	0.58	Moderate
Spleen	4	3	6	24	57.14%	80.00%	75.68%	0.07*	0.32	Fair
Stomach	3	2	7	25	60.00%	78.13%	75.68%	0.11*	0.27	Fair
Female mice										
Bone Marrow	0	1	1	35	0.00%	97.22%	94.59%	1.00*	−0.04	No
Kidney	1	2	6	28	11.11%	82.35%	78.38%	0.48*	−0.59	No
Liver	11	2	5	19	61.11%	79.17%	81.08%	0.00	0.54	Moderate
Lung	1	0	3	33	25.00%	91.67%	91.89%	0.11*	−0.61	No
Nose	5	2	3	27	50.00%	90.00%	86.49%	0.00*	0.54	Moderate
Spleen	4	3	4	26	36.36%	86.67%	81.08%	0.03*	0.34	Fair
Stomach	2	3	6	26	18.18%	81.25%	75.68%	0.29*	−0.15	No
Male mice										
Bone Marrow	0	2	1	34	0.00%	97.14%	91.89%	1.00*	0.20	Small
Kidney	2	1	5	29	25.00%	85.29%	83.78%	0.09*	−0.17	No
Liver	12	1	5	19	66.67%	79.17%	83.78%	0.00	0.60	Substantial
Lung	1	0	2	34	33.33%	94.44%	94.59%	0.08*	−0.06	No
Nose	5	2	3	27	50.00%	90.00%	86.49%	0.00*	0.54	Moderate
Spleen	4	3	4	26	36.36%	86.67%	81.08%	0.03*	0.34	Fair
Stomach	3	2	5	27	30.00%	84.38%	81.08%	0.06*	0.12	Small
Rats										
Bone Marrow	3	5	2	27	37.50%	93.10%	81.08%	0.06*	0.35	Fair
Kidney	10	5	4	18	66.67%	81.82%	75.68%	0.01	0.49	Moderate
Liver	17	2	6	12	89.47%	66.67%	78.38%	0.00	0.56	Moderate
Lung	1	2	3	31	33.33%	91.18%	86.49%	0.30*	0.21	Fair
Nose	6	1	2	28	85.71%	93.33%	91.89%	0.00*	0.75	Substantial
Spleen	7	2	2	26	77.78%	92.86%	89.19%	0.00*	0.71	Substantial
Stomach	4	5	4	24	44.44%	85.71%	75.68%	0.08*	0.31	Fair
Female rats										
Bone Marrow	3	5	1	28	37.50%	96.55%	83.78%	0.03*	0.42	Moderate
Kidney	7	6	3	21	43.75%	87.50%	75.68%	0.02*	0.43	Moderate
Liver	15	3	6	13	62.50%	68.42%	75.68%	0.00	0.42	Moderate
Lung	1	2	3	31	16.67%	91.18%	86.49%	0.30*	0.05	Small
Nose	5	2	2	28	55.56%	93.33%	89.19%	0.00*	0.64	Substantial
Spleen	6	2	2	27	60.00%	93.10%	89.19%	0.00*	0.67	Substantial
Stomach	2	6	3	26	18.18%	89.66%	75.68%	0.29*	0.24	Fair
Male rats										
Bone marrow	2	5	1	29	25.00%	96.67%	83.78%	0.09*	0.46	Moderate
Kidney	9	5	4	19	50.00%	82.61%	75.68%	0.01*	0.43	Moderate
Liver	16	1	6	14	69.57%	70.00%	81.08%	0.00	0.55	Moderate
Lung	1	1	1	34	33.33%	97.14%	94.59%	0.11*	0.46	Moderate
Nose	6	1	2	28	66.67%	93.33%	91.89%	0.00*	0.73	Substantial
Spleen	6	2	1	28	66.67%	96.55%	91.89%	0.00*	0.76	Substantial
Stomach	4	3	3	27	40.00%	90.00%	83.78%	0.01*	0.44	Moderate

Notes: The *p* values were computed from either Fisher's exact tests or Chi-squared tests for independence. Fisher's exact was used if any cell contained fewer than five chemicals. Entries with (*) indicate *p* values were computed from Fisher's exact test; otherwise, the Chi-square test was used.

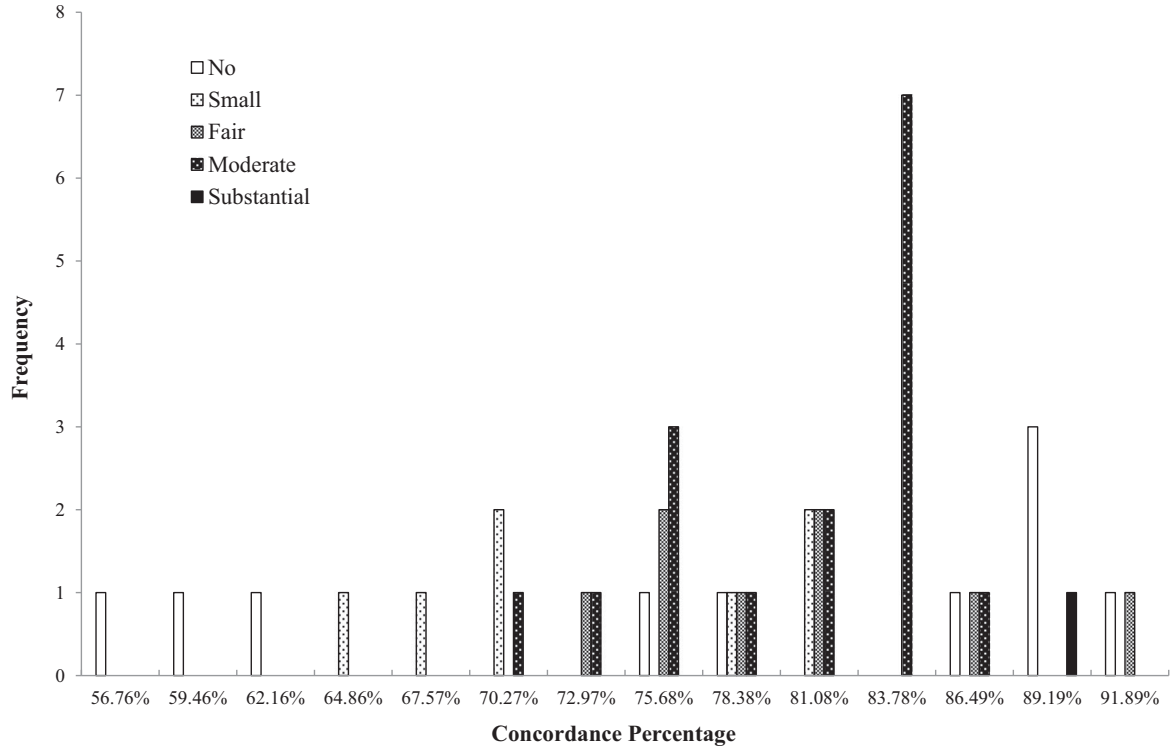

Fig. 2. Frequency distribution of interspecies concordance percentages shown in Table IV ($N = 42$) by Kappa value agreement category.

both females and males. The difference was due to a larger number of chemicals producing kidney lesions in rats than mice. For example, in females after short-term exposure, three chemicals produced lesions in mouse kidney and two of the three also produced kidney lesions in rats (mouse-to-rat PPV = 33.3%), while 13 chemicals produced lesions in rat kidney and only one also produced kidney lesions in mice (rat-to-mouse PPV = 7.7%).

NPV values were much higher across the board, with most values in the 80% and 90% range, which indicated that the absence of lesions in a specific organ in mice was usually associated with no lesions in that organ in rats, and *vice versa*.

Overall, lesions observed in mouse liver and nose were most likely to be observed in the same organ in rats, and *vice versa*. For kidney and bone marrow lesions, mice showed a higher predictive ability to rats, than rats to mice. A higher interspecies concordance was shown after longer chemical exposure in males. Therefore, in the combination organ, exposure length and sex, the highest PPV observed was 88.2% from mice to rats for liver lesions in males after a long-term exposure. All other PPVs mostly center around 50%, falling in the range of 30% to 70%. NPV (lack of lesions in a specific organ in mice predicting no response in rats, and *vice versa*) was relatively high.

3.2. Prediction from Short-Term to Long-Term Tests

Like interspecies nondirectional concordance, the nondirectional concordance in lesions between short-term and long-term exposures was measured by concordance percentage, Kappa value, Kappa agreement category, and independence tests. The question is whether lesions in a given organ in the short-term exposure length would be more likely to occur in the same organ after a long-term exposure to the same chemical and whether there would be a difference in agreement between rats and mice. The interexposure length concordance percentages were all relatively high, ranging from around 70% to almost 100% (Table V). Kappa values ranged from −0.06 (poor agreement) to 0.62 (substantial agreement) with an average of 0.32 (fair agreement) in mice and from 0.21 (fair agreement) to 0.75 (substantial agreement) with an average of 0.48 (moderate agreement) in rats. As shown in Tables IV and V, interspecies

Kappa agreement categories were more frequently in the range of slight to moderate, while interexposure length agreement categories were more frequently in the range of fair to substantial. In mice, the organ with the highest interexposure length Kappa value was liver, followed by nose, spleen, lung, stomach, kidney, and bone marrow. In rats, the order was nose, spleen, liver, kidney, bone marrow, stomach, and lung. Higher interexposure length Kappa values were observed in males than in females. For example, the mean Kappa values across organs in male rodents, male mice, and male rats were 0.48, 0.22, and 0.55, while they were 0.38, 0.003, and 0.41 in female rodents, mice, and rats.

To examine whether the lesions found in long-term bioassays were predicted by the shorter term studies, short-to-long PPV and NPV for the same sex/species were computed and interpreted. In mice, short-to-long PPVs were highest (100%) for lung, followed by liver (92.3%), nose (71.4%), kidney (66.7%), stomach (60%), spleen (57%), and a low 0% for bone marrow. In rats, the short-to-long PPVs were similar to those in mice for kidney, liver, nose, spleen, and stomach, but higher for bone marrow (37.5% in rats vs. 0% in mice) and lower for lung (33.3% in rats vs. 100% for mouse lung). Short-to-long PPVs in male mice were consistently similar to or higher than those in female mice. Similar differences in short-to-long PPVs were observed between male and female rats. NPV values were much higher across the board, which indicated lesions in a specific organ not observed in short-term studies would probably not be observed in long-term studies.

Compared to the interspecies nondirectional concordance of lesions responding to a chemical exposure, short-term to long-term concordance is higher. This finding is also supported by result of the Chi-square or Fisher's exact tests. It can be seen in Tables IV and V that for interspecies concordance, 24 (38%) p values were less than 0.05, indicating organ-specific nonneoplastic lesions were related between mice and rats. For interexposure length concordance, more p values were less than 0.05 (35, 56%), indicating organ-specific nonneoplastic lesions were related between short- and long-term exposures.

4. DISCUSSION

The question of prediction of adverse effects across species—how well the animal models used in toxicology might predict human responses to the same exposure—has important implications for toxicity testing, for risk assessment, and for policy. It is often assumed that toxicity tests of many systems and functions are needed to characterize the adverse effects associated with a chemical.[10] If effects are poorly concordant across even rodent species it would suggest that identification of specific effects likely to occur in humans would be very difficult (lacking specific mode of action or toxicity pathway data).

To investigate this question we looked at both cross-species and cross-duration concordance and predictive value for noncancer lesions (aggregated at the organ level) in rodent chemical bioassays. If we think of each experiment (male and female mice and rats) as replications of a test for the induction of specific pathologic lesions, then concordance tells us about the reproducibility of the lesion in different tests. Of course, the key question, which cannot be addressed here, is the concordance between test rodents and humans. Our use of Kappa essentially measures the degree of agreement about each lesion by treating each experiment as a separate observation of the biological response to the chemical.

By all measures, but even more so using Kappa, the concordance of response across species/sex combinations is far from perfect in both long-term and short-term studies. This is a bit surprising given the high doses of chemical given to these phylogenetically similar animals under tightly controlled conditions. Clearly, there are sex- and species-specific modes of toxicity that have been worked out,[11] but these results suggest that, absent some *a priori* chemical-specific knowledge, prediction of a specific noncancer pathologic lesion, even from mice to rats, is quite imprecise. Concordance of toxicologic lesions is much higher between short-term and long-term tests of the same chemical. It is a bit surprising that there are a significant fraction of tissues that develop pathologies during short-term exposure that are not identified as compromised following long-term exposure (e.g., for male rats five chemicals were identified as having kidney lesions in short-term tests but not in long-term ones) and *vice versa*.

We calculated predictive values (both PPV and NPV) to address two key questions:

(1) If I observe pathology in an organ in one species (or sex or assay length), how likely is it I will observe the same pathology in another species exposed to the chemical (and in this case under identical circumstances)?

(2) If I DON'T observe pathology in an organ in one species, how likely is it I will NOT observe the same pathology in another species exposed to the chemical (and in this case under identical circumstances)?

Both of these are important questions for the use of animal data in predicting human risk. In addressing question 1 we find the highest value of PPV between mice and rats for lesions of the liver. In short-term studies, 85% of the times a liver lesion was seen in mice it was also seen in rats. In long-term studies the corresponding value is 89%. All other PPVs are less than 60% (Table IV). Predicting within a sex (FM to FR, MM to MR) yielded similar values for both short- and long-term studies. Other than bone marrow, where female mice predicted female rats perfectly (although not *vice versa*), the liver was again the organ with highest PPV and values were, in general, higher than the prediction when the sexes were pooled. For all but two comparisons (females, long-term nose and stomach) PPV was higher from mice to rats than from rats to mice. Overall, in response to question 1, it appears that if a pathology is observed in an organ in one species (or sex or assay length), it is not very likely, on average, that the same pathology will be observed in another species exposed to the chemical.

NPV helps evaluate question 2. NPV values both across species and within a sex across species were higher than PPVs. Much of this is due to the large number of observations in the −/− cell, that is, when a lesion was not found in either group being compared. Overall NPVs ranged from about 62% to 100% and it does not appear that there is a discernable difference between the predictions from mice to rats versus rats to mice. From these results it appears that the absence of a pathologic response in a specific tissue is a good predictor of that lesion not occurring in another group (sex or species) exposed to the same chemical.

To the best of our knowledge, this is the first study evaluating the concordance of noncancer lesions across species. There have been evaluations of the concordance of carcinogenic lesions in long-term rodent bioassays,[2,12] that which generally find moderate concordance (in the range of 70–80%) for carcinogenesis overall but very poor concordance of the actual tumor types. This study has several strengths for addressing our question, especially our use of data from NTP-sponsored studies. These are very well-conducted and controlled experiments using virtually identical conditions, including pathology. Another strength is the use of the NTP pathologists' judgments about which lesions are significant. Finally, we use specific and consistent inclusion criteria for the bioassays chosen for study.

There are limitations to this study, due primarily to the relatively small number of chemicals compared. Although the NTP endeavors to make the results of bioassays available electronically, the specific endpoints we wished to evaluate had to be abstracted from NTP Technical Reports. Also potentially important are possible changes in NTP practices or pathology nomenclature over time. Rolling up lesions to the level of organ was one way to address this, although it does lose some specificity. Also, studies being compared were done over a relatively short span of time so changes over time may not be a major concern.

For many uses of risk information, the ability to predict human outcomes from toxicologic data is crucial. However, this study and others[1,2,12] suggest the specificity of predictions is lacking. Serious questions about the applicability of the standard rodent bioassay have been raised.[13] It is possible that new approaches to toxicological testing based on adverse outcome pathways will help[14,15] as these tools are developed. On the other hand, the ability to explain (and predict) nonconcordance of response will be a challenge to the development of adverse outcome pathways. All will need to be considered in a weight-of-evidence approach to causation of specific noncancer effects.[16] While we are waiting for this new paradigm to develop, however, risk assessment will continue to use traditional toxicology tests like those relied upon for this study.

It may be that difficulty in predicting the actual outcome of an exposure will lead risk assessment to focus on the exposure or dose required to elicit some adverse effect without specification of the effect.[17] However, this would suggest that risk assessment can identify exposures at which adverse effects might occur but not predict the actual outcomes of exposure. The inability to predict outcomes, necessary for either benefit-cost analysis or cost-effectiveness analysis, makes risk assessment less useful for policy analysis.

ACKNOWLEDGMENTS

The authors with to thank Nada Raooff, MPH, for excellent technical assistance.

REFERENCES

1. Olson H, Betton G, Robinson D, Thomas K, Monro A, Kolaja G, G., Lilly P, Sanders J, Sipes G, Bracken W, Dorato M, Van Deun K, Smith P, Berger B, Heller A. Concordance of the toxicity of pharmaceuticals in humans and in animals. Regulatory Toxicology and Pharmacology, 2000; 32(1):56–67.
2. Gold LS, Bernstein L, Magaw R, Slone TH. Interspecies extrapolation in carcinogenesis: Prediction between rats and mice. Environmental Health Perspectives, 1989; 81:211–219.
3. Linkov I, Wilson R, Gray GM. Anticarcinogenic responses in rodent cancer bioassays are not explained by random effects. Toxicological Sciences, 1998; 43(1):1–9.
4. Chhabra R, Huff J, Schwetz B, Selkirk J. An overview of prechronic and chronic toxicity/carcinogenicity experimental study designs and criteria used by the national toxicology program. Environmental Health Perspectives, 1990; 86:313–321.
5. Bucher JR. The national toxicology program rodent bioassay. Annals of the New York Academy of Sciences, 2002; 982(1):198–207.
6. National Toxicology Program. Pathology code table: Microscopic terms by organ (large report), 2013.
7. Viera AJ, Garrett JM. Understanding interobserver agreement: The kappa statistic. Family Medicine, 2005; 37(5):360–363.
8. Carletta J. Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics, 1996; 22(2):249–254.
9. Cochran WG. Some methods for strengthening the common [chi-squared] tests. Biometrics, 1954; 10:417–451.
10. Holman E, Gray G. Setting pesticide reference doses: A retrospective analysis examining key data and choices. Human and Ecological Risk Assessment: An International Journal, 2013; 20:1550–1564.
11. Swenberg JA. Alpha 2u-globulin nephropathy: Review of the cellular and molecular mechanisms involved and their implications for human risk assessment. Environmental Health Perspectives, 1993; 101(Suppl 6):39–44.
12. Haseman JK, Lockhart AM. Correlations between chemically related site-specific carcinogenic effects in long-term studies in rats and mice. Environmental Health Perspectives, 1993; 101(1):50–54.
13. Marone PA, Hall WC, Hayes AW. Reassessing the two-year rodent carcinogenicity bioassay: A review of the applicability to human risk and current perspectives. Regulatory Toxicology and Pharmacology, 2014; 68(1):108–118.
14. Collins FS, Gray GM, Bucher JR. Toxicology. Transforming environmental health protection. Science, 2008; 319(5865):906–907.
15. Tice RR, Austin CP, Kavlock RJ, Bucher JR. Improving the human hazard characterization of chemicals: A Tox21 update. Environmental Health Perspectives, 2013; 121(7):756–765.
16. Rhomberg L. Hypothesis-based weight of evidence: An approach to assessing causation and its application to regulatory toxicology. Risk Analysis, 2014. DOI: 10.1111/risa.12206.
17. Thomas RS, Philbert MA, Auerbach SS, Wetmore BA, Devito MJ, Cote I, Rowlands JC, Whelan MP, Hays SM, Andersen ME, Meek ME, Reiter LW, Lambert JC, Clewell HJ III, Stephens ML, Zhao QJ, Wesselkamper SC, Flowers L, Carney EW, Pastoor TP, Petersen DD, Yauk CL, Nong A. Incorporating new technologies into toxicity testing and risk assessment: Moving from 21st century vision to a data-driven framework. Toxicological Sciences, 2013; 136(1):4–18.