# Data Quality in Predictive Toxicology: Reproducibility of Rodent Carcinogenicity Experiments

*Eva Gottmann,[1,2] Stefan Kramer,[3] Bernhard Pfahringer,[4] and Christoph Helma[1,2,3]*

[1]Institute for Cancer Research, and [2]Institute for Environmental Hygiene, University Vienna, Vienna, Austria; [3]Institute for Computer Science, Machine Learning Lab, University Freiburg, Freiburg, Germany; [4]Austrian Research Institute for Artificial Intelligence, Vienna, Austria

We compared 121 replicate rodent carcinogenicity assays from the two parts (National Cancer Institute/National Toxicology Program and literature) of the Carcinogenic Potency Database (CPDB) to estimate the reliability of these experiments. We estimated a concordance of 57% between the overall rodent carcinogenicity classifications from both sources. This value did not improve substantially when additional biologic information (species, sex, strain, target organs) was considered. These results indicate that rodent carcinogenicity assays are much less reproducible than previously expected, an effect that should be considered in the development of structure–activity relationship models and the risk assessment process. *Key words*: carcinogenicity, machine learning, predictive toxicology, quality assurance, structure–activity relationships. *Environ Health Perspect* 109:509–514 (2001). [Online 9 May 2001]
*http://ehpnet1.niehs.nih.gov/docs/2001/109p509-514gottmann/abstract.html*

The development of structure–activity relationships (SARs) is gaining more and more importance in predictive toxicology and risk assessment. These models rely on the comparison of chemical structures and their properties with their toxicologic effects and can be used for the prediction of adverse effects of chemicals, but they are also valuable tools to investigate questions of scientific interest (e.g., toxicologic mechanisms). Each SAR study needs reliable chemical and biologic data, but this aspect is neglected in most investigations. Few systemic studies are available for the development of SAR models. This article presents a discussion about the reliability of toxicologic data in SAR models and risk assessment. In a previous paper (*1*), we covered the identification and representation of chemical structures and the calculation of chemical properties.

The database we used for our investigation was the Carcinogenic Potency Database (CPDB) (*2*). It contains detailed information from long-term *in vivo* carcinogenicity experiments for 1,289 structurally diverse (noncongeneric) compounds. It consists of two major parts. One data set contains the results of carcinogenicity experiments performed by the National Cancer Institute (NCI) and the National Toxicology Program (NTP); the other part contains carcinogenicity assays from the general literature that meet certain quality criteria (e.g., concerning administration of compounds, duration of experiments, number of test and control animals, availability of original data). One hundred twenty-one chemicals were tested in both parts [Table 1; the complete data set is available from the authors (*3*)]. This overlap allows the comparison of replicate carcinogenicity experiments with the same compound.

Our intention was to investigate the reliability of rodent carcinogenicity assays for SAR studies and risk assessment. This is usually ascertained by repeating experiments with the same substance under the same test conditions. Rodent carcinogenicity experiments, however, are too time consuming and expensive to replicate experiments for a sufficiently large number of compounds. Therefore, we compared carcinogenicity experiments of the overlapping compounds from the NCI/NTP and the literature parts, although they were not performed with identical protocols. This closely resembles the real-world situation for the development of SAR models, where results from different sources and different protocols are combined to obtain a larger database.

One measure of reproducibility is concordance, the number or percentage of chemicals that are classified the same way in different data sets. More precisely, we prepared first an overall rodent carcinogenicity classification because classifications neglecting additional information such as species, sex, and strain are often used as the basis for SAR studies and risk assessment. In further comparisons we considered species- and sex-specific effects because it was recommended to use these parameters in SAR models (*4*) and for the registration of chemicals. We also intended to find out whether it is sensible to develop organ-specific SARs and to what extent additional biologic information influences the accuracy of SAR models and risk assessment. We investigated whether additional toxicologic information, in our case mutagenicity, can help identify carcinogens and whether it should be included in SAR studies. Finally, we compared the quantitative measure for carcinogenic potency, the tumorigenic dose rate 50 ($TD_{50}$) (*5,6*) to find out if it is sensible to predict $TD_{50}$ values and to distinguish between strong and weak carcinogens.

## Methods

*Source of data.* Toxicologic data were obtained from the CPDB compiled by Gold et al. (*2*). This is the most extensive and detailed publicly available carcinogenicity database, with results of chronic, long-term animal cancer tests. Both qualitative and quantitative information on positive and negative experiments are given, including all bioassays from the NCI/NTP and results from the general literature that meet a set of inclusion criteria (*2*). The CPDB contains experiments with 1,298 chemicals. For each experiment, the following information is included: species, strain, and sex of test animals; features of the experimental protocol such as route of administration, duration of dosing, dose levels in milligrams per kilogram of body weight per day, duration of the experiment; histopathology and tumor incidence; carcinogenic potency ($TD_{50}$) and its statistical significance; shape of the dose–response curve; authors' opinion as to carcinogenicity; and literature citation (*2*). The CPDB data were converted to Prolog facts for Machine Learning Experiments and data analysis within Sicstus Prolog (*7*).

*Statistical evaluation.* Generally the results of our investigations take the form shown in Table 2, which summarizes concordant and discordant classifications (carcinogen/noncarcinogen) from two different data sources (NCI/NTP, literature). To analyze the degree to which classifications of items are associated, we used the G index according to Holley and Guilford (*8*) and the association coefficient according to Cole (*9*). Given such

Address correspondence to E. Gottmann, Federal Ministry of Education, Science and Culture, Abt. VII/B/5 - PROVISO, Minoritenplatz 5, A-1014 Vienna Austria. Telephone: +43-1-53120-5850. Fax: +43-1-53120-5805. E-mail: eva.gottmann@bmbwk.gv.at

We thank M. Kundi for helpful discussions and suggestions for improvements. This project was funded by the Austrian Federal Ministry of Science and Transport (GZ 70.017/2-Pr/4/87). Partial support was also provided by the "Jubiläumsfond der Österreichischen Nationalbank" under grant 6930. The Austrian Federal Ministry of Science and Transport provides general financial support for the Austrian Research Institute for Artificial Intelligence.

Received 18 August 2000; accepted 6 December 2000.

a table consisting of four entries, *a*, *b*, *c*, and *d*, as defined in Table 2 (inadequate results were not considered in statistical analysis), these statistics are defined as follows.

The G index is defined as

$$G = \frac{(a+d)-(b+c)}{N}$$

It is a measure of the relative increase or decrease in the number of concordant classifications. The assumption of independence ($H_0$: $G = 0$) is tested using the sign test, either exactly or using the normal distribution:

$$z = \frac{2(b+c)-N}{\sqrt{N}}$$

Given positive association, the association coefficient according to Cole (9) is defined as follows:

$$C = \frac{ad - bc}{\min(b,c)N + ad \text{ - } bc}$$

The asymptotic test for $C = 0$ is

**Table 1.** Identification of the 121 compounds tested in the NTP and literature part of the CPDB.

| Abbreviation | Chemical | CAS No. | Abbreviation | Chemical | CAS No. |
|---|---|---|---|---|---|
| 1te | 1,1,1-Trichloroethane, technical grade | 71-55-6 | kep | Kepone | 143-50-0 |
| 2dc | 1,2-Dichloroethane | 107-06-2 | las | Lasiocarpine | 303-34-4 |
| 3db | 3,3'-Dimethoxybenzidine · 2HCl | 20325-40-0 | ldt | Lead dimethyldithiocarbamate | 19010-66-3 |
| ald | Aldrin | 309-00-2 | lin | γ-1,2,3,4,5,6-Hexachlorocyclohexane | 58-89-9 |
| ant | 2-Amino-5-nitrothiazole | 121-66-4 | mbr | Methyl bromide | 74-83-9 |
| apc | Aspirin, phenacetin, and caffeine | 8003-03-0 | mbt | 2-Mercaptobenzothiazole | 149-30-4 |
| azb | Azobenzene | 103-33-3 | mca | Monochloroacetic acid | 79-11-8 |
| azc | 5-Azacytidine | 320-67-2 | mcl | Mercuric chloride | 7487-94-7 |
| b38 | C.I. Direct black 38 | 1937-37-7 | mop | 8-Methoxypsoralen | 298-81-7 |
| bcm | Bromodichloromethane | 75-27-4 | mrx | Mirex | 2385-85-5 |
| bde | 1,3-Butadiene | 106-99-0 | mxc | Methoxychlor | 72-43-5 |
| ben | Benzene | 71-43-2 | myc | Methylene chloride | 75-09-2 |
| bht | Butylated hydroxytoluene | 128-37-0 | nac | 5-Nitroacenaphthene | 602-87-9 |
| bna | Benzyl acetate | 140-11-4 | nat | Nitrilotriacetic acid, trisodium salt, monohydrate | 18662-53-8 |
| bzo | Coumarin | 91-64-5 | nff | 1-[(5-Nitrofurfurylidene)amino]hydantoin | 67-20-9 |
| cap | Captan | 133-06-2 | nha | 3-Nitro-4-hydroxyphenylarsonic acid | 121-19-7 |
| ccc | (2-Chloroethyl)trimethylammonium chloride | 999-81-5 | nta | Nitrilotriacetic acid | 139-13-9 |
| cci | Cyanamide, calcium | 156-62-7 | oca | Ochratoxin A | 303-47-9 |
| cdu | 3-(p-Chlorophenyl)-1,1-dimethylurea | 150-68-5 | pb1 | Aroclor 1254 | 11097-69-1 |
| chb | Chlorobenzilate | 510-15-6 | pbt | Phenylbutazone | 50-33-9 |
| chd | Chlordane, technical grade | 57-74-9 | pch | 2,3,4,5,6-Pentachlorophenol (Dowicide EC-7) | 87-86-5 |
| chf | Chloroform | 67-66-3 | pcm | Picloram, Technical grade | 1918-02-1 |
| cms | C.I. Food Red 3 | 3567-69-9 | pct | 2,3,4,5,6-Pentachlorophenol, technical grade | 87-86-5 |
| cpm | Chlorpheniramine maleate | 113-92-8 | pdd | *p,p'*-DDD | 72-54-8 |
| ctl | 4-Chloro-*o*-toluidine · HCl | 3165-93-3 | pde | *p,p'*-DDE | 72-55-9 |
| dan | 2,4-Diaminoanisole sulfate | 39156-41-7 | phn | 1-Phenylazo-2-naphthol | 842-07-9 |
| day | C.I. Pigment yellow 12 | 6358-85-6 | pip | Piperonyl butoxide | 51-03-6 |
| dbe | 1,2-Dibromoethane | 106-93-4 | pna | Phenyl-β-naphthylamine | 135-88-6 |
| dcv | Dichlorvos | 62-73-7 | pnb | Pentachloronitrobenzene | 82-68-8 |
| ddt | DDT | 50-29-3 | pni | *p*-Nitroaniline | 100-01-6 |
| deu | *N,N'*-Diethylthiourea | 105-55-5 | prb | Procarbazine · HCl | 366-70-1 |
| dhm | Diphenhydramine · HCl | 147-24-0 | prg | Propyl gallate | 121-79-9 |
| dhx | Di(2-ethylhexyl)phthalate | 117-81-7 | prl | Propylene | 115-07-1 |
| die | Dieldrin | 60-57-1 | prp | 1,2-Propylene oxide | 75-56-9 |
| dio | 1,4-Dioxane | 123-91-1 | psu | Piperonyl sulfoxide | 120-62-7 |
| dis | Tetraethylthiuram disulfide | 97-77-8 | qrc | Quercetin | 117-39-5 |
| dmz | 3,3'-Dimethylbenzidine · 2HCl | 612-82-8 | red | *N*-Nitrosodiphenylamine | 86-30-6 |
| dph | 5,5-Diphenylhydantoin | 57-41-0 | ros | *p*-Rosaniline · HCl | 569-61-9 |
| dr9 | D & C Red No. 9 | 5160-02-1 | rsp | Reserpine | 50-55-5 |
| dsa | Daminozide | 1596-84-5 | saz | Azide, sodium | 26628-22-8 |
| edd | *p,p*'-Ethyl-DDD | 72-56-0 | sdc | Sodium diethyldithiocarbamate trihydrate | 148-18-5 |
| egl | Eugenol | 97-53-0 | sma | Malonaldehyde, sodium salt | 24382-04-5 |
| ela | Ethyl acrylate | 140-88-5 | sof | Fluoride, sodium | 7681-49-4 |
| end | Endrin | 72-20-8 | sta | Tin (II) chloride | 7772-99-8 |
| eod | Ethylene oxide | 75-21-8 | sty | Styrene | 100-42-5 |
| eta | Ethionamide | 536-33-4 | tcb | 2,4,6-Trichlorophenol | 88-06-2 |
| ete | Ethyl tellurac | 20941-65-5 | tcd | 2,3,7,8-Tetrachlorodibenzo-*p*-dioxin | 1746-01-6 |
| eth | Ethylene thiourea | 96-45-7 | tce | Trichloroethylene | 79-01-6 |
| ffl | Furfural | 98-01-1 | tda | 4,4'-Thiodianiline | 139-65-1 |
| fl2 | Trichlorofluoromethane | 75-69-4 | tep | THIO-TEPA | 52-24-4 |
| fsz | 5-Nitro-2-furaldehyde semicarbazone | 59-87-0 | thd | Endosulfan | 115-29-7 |
| fy6 | FD & C Yellow No. 6 | 2783-94-0 | tol | Toluene | 108-88-3 |
| gar | Tetrachlorvinphos | 961-11-5 | tou | *o*-Toluidine · HCl | 636-21-5 |
| gly | Glycidol | 556-52-5 | trf | Trifluralin, technical grade | 1582-09-8 |
| hb1 | HC Blue No. 1 | 2784-94-3 | trs | Tris(2,3-dibromopropyl)phosphate | 126-72-7 |
| hcp | Hexachlorophene | 70-30-4 | try | L-Tryptophan | 73-22-3 |
| hct | Hydrochlorothiazide | 58-93-5 | tub | Rotenone | 83-79-4 |
| hep | Heptachlor | 76-44-8 | vdc | Vinylidene chloride | 75-35-4 |
| hql | 8-Hydroxyquinoline | 148-24-3 | zdd | Zinc dimethyldithiocarbamate | 137-30-4 |
| hya | Acetaminophen | 103-90-2 | zec | Mexacarbate | 315-18-4 |
| hyq | Hydroquinone | 123-31-9 | | | |

Results from carcinogenicity assays are available from the authors (*3*).

$$z = \frac{c}{\sqrt{\frac{(a+c)(c+d)}{N(a+b)(b+d)}}}$$

Both indices may range from 0 to 1, where 1 indicates an ideal association. In Table 3, we summarize the results for these statistical tests, include an interpretation (using the standard interpretation scale for the normal distribution), and rank the results accordingly (Table 2).

## Results

As described above, the CPDB consists of two different subsets, the results from the NCI/NTP and the results from the general literature. From experiments on 1,298 chemical agents, only 121 chemicals were tested in both the NCI/NTP and the literature parts. We used this overlapping part of the CPDB for the present analysis.

*Overall rodent carcinogenicity classification.* Our aim was to quantify the concordance/discordance between carcinogenicity classifications from both data sets. Classifications were based on authors' opinions because authors consider more than statistical significance alone [historical control rates for particular sites, survival and latency, and/or dose response (*2*)]. A compound was classified as a carcinogen if a positive result was obtained in at least one experiment.

The results are summarized in Table 4, which shows that an unexpectedly low proportion of these 121 compounds were classified concordantly in both parts as carcinogens or noncarcinogens. Only 69 (57%) chemicals had concordant authors' opinions, 57% (39/69) of the chemicals were consistently classified as positive, and 43% (30/69) had negative results in both sources (Table 4).

*Species- and sex-specific effects.* We studied the detail, considering the reproducibility of species and sex specific effects. Evans et al. (*10*) already noted that species- and sex-specific tumorigenicity is one of the parameters that should be considered when evaluating results from animal cancer studies. Our calculations gave the following results (Table 5): from 70 investigations with mice, 49% had concordant results; from 71 experiments with rats, 62% were concordant. The consideration of sexes gave similar results. The concordance of male mice was 46%, of female mice 36%, of male rats 55%, and of female rats 69% (Table 5).

**Table 2.** Definition of variables for the statistical tests.

| | Carcinogen[a] | Noncarcinogen[b] | Literature |
|---|---|---|---|
| Carcinogen[a] | *a* | *b* | *a* + *b* |
| Noncarcinogen[b] | *c* | *d* | *c* + *d* |
| NCI/NTP | *a* + *c* | *b* + *d* | *N* = *a* + *b* + *c* + *d* |

[a]At least one experiment is evaluated as positive. [b]At least one experiment is evaluated as negative and no experiment is evaluated as positive.

**Table 3.** Overview of the statistical evaluation.

| | *G* index | *z* | Int[a] | Cole | *z* | Int | Rank |
|---|---|---|---|---|---|---|---|
| Overall | 0.353 | –3.565 | Very high | 0.407 | 3.578 | Very high | 3 |
| Mice | 0.283 | –2.060 | High | 0.351 | 3.095 | Very high | 6 |
| Rats | 0.375 | –3.000 | Very high | 0.398 | 2.991 | Very high | 3 |
| Female mice | 0.200 | –1.264 | Low | 0.205 | 1.433 | Low | 7 |
| Male mice | 0.333 | –2.160 | High | 0.467 | 4.058 | Very high | 5 |
| Female rats | 0.522 | –3.539 | Very high | 0.489 | 3.481 | Very high | 1 |
| Male rats | 0.373 | –2.661 | Very high | 0.440 | 2.678 | Very high | 2 |
| Single sp./multiple sp.[b] | 0.152 | –0.870 | Low | 0.313 | 3.022 | Very high | 8 |
| Single cat./multiple cat.[c] | 0.030 | –0.174 | Low | — | — | — | 9 |

[a]Interpretation according to the standard scale for the normal distribution.[b]Single-species/multiple-species carcinogens. [c]Single-category/multiple-categories carcinogens.

**Table 4.** A comparison of the classification in the NCI/NTP and literature parts of the CPDB.

| | Carcinogen[a] | Noncarcinogen[b] | Inadequate[c] | Literature |
|---|---|---|---|---|
| Carcinogen[a] | 39 | 13 | 1 | 53 |
| Noncarcinogen[b] | 20 | 30 | 0 | 50 |
| Inadequate[c] | 10 | 8 | 0 | 18 |
| NCI/NTP | 69 | 51 | 1 | 121 |

Concordant classification: 69 compounds (57%); discordant classification: 52 compounds (43%).
[a]At least one experiment is evaluated as positive. [b]At least one experiment is evaluated as negative and no experiment is evaluated as positive. [c]Experiments are evaluated neither positive nor negative.

**Table 5.** A comparison of the classification with consideration on species and sex in the NCI/NTP and literature parts of the CPDB.

| | Mice | | | | Rats | | | |
|---|---|---|---|---|---|---|---|---|
| | Carcinogenic[a] | Noncarcinogenic[b] | Inadequate[c] | Literature | Carcinogenic[a] | Noncarcinogenic[b] | Inadequate[c] | Literature |
| Males | | | | | | | | |
| Carcinogenic[a] | 11 | 10 | 1 | 22 | 15 | 6 | 1 | 22 |
| Noncarcinogenic[b] | 4 | 17 | 1 | 22 | 10 | 20 | 3 | 33 |
| Inadequate[c] | 4 | 12 | 1 | 17 | 4 | 5 | 0 | 9 |
| NCI/NTP | 19 | 39 | 3 | 61 | 29 | 31 | 4 | 64 |
| Concordant | 28 compounds (46%) | | | | 35 compounds (55%) | | | |
| Discordant | 33 compounds (54%) | | | | 29 compounds (45%) | | | |
| Females | | | | | | | | |
| Carcinogenic[a] | 9 | 9 | 1 | 19 | 10 | 6 | 1 | 17 |
| Noncarcinogenic[b] | 7 | 15 | 3 | 25 | 5 | 25 | 1 | 31 |
| Inadequate[c] | 6 | 13 | 3 | 22 | 1 | 1 | 1 | 3 |
| NCI/NTP | 22 | 37 | 7 | 66 | 16 | 32 | 3 | 51 |
| Concordant | 24 compounds (36%) | | | | 35 compounds (69%) | | | |
| Discordant | 42 compounds (64%) | | | | 16 compounds (31%) | | | |
| Both | | | | | | | | |
| Carcinogenic[a] | 15 | 12 | 0 | 27 | 20 | 9 | 1 | 30 |
| Noncarcinogenic[b] | 7 | 19 | 0 | 26 | 11 | 24 | 0 | 35 |
| Inadequate[c] | 4 | 12 | 1 | 17 | 4 | 2 | 0 | 6 |
| NCI/NTP | 26 | 43 | 1 | 70 | 35 | 35 | 1 | 71 |
| Concordant | 34 compounds (49%) | | | | 44 compounds (62%) | | | |
| Discordant | 36 compounds (51%) | | | | 27 compounds (38%) | | | |

[a]At least one experiment is evaluated as positive. [b]At least one experiment is evaluated as negative, and no experiment is evaluated as positive. [c]Experiments are evaluated as neither positive nor negative.

There is obviously a difference between the reproducibility of experiments with mice and rats, although mice and rats are closely related. The tests on rats seem to be far more reproducible. Concerning the sexes, we could not observe clear trends. On one hand, the reproducibility of experiments on male mice was better than on female mice; on the other hand, experiments on female rats had a better concordance than those on male rats. It is notable that experiments with female mice are, with a high probability, statistically independent (Table 3).

Carcinogens in mice are probably heavily influenced by the strain (*3*), which shows that the selection of the strains is another important source of variability in the data. In the NCI/NTP studies, 3 different rat strains and 1 mouse strain were used for experiments with the 121 overlapping chemicals; in the literature, 29 different rat strains and 37 different mice strains were tested. The most frequently used strains in both parts of the CPDB were Fischer F344/N rats and B6C3F$_1$ mice. The concordance for these strains was 53% (9/17) for male rats, 64% (7/11) for female rats, 39% (15/38) for male mice, and 33% (13/40) for female mice. These values are close to the overall concordance (Table 5) and indicate that the poor reproducibility of carcinogenicity assays may not be due to different strains in the NCI/NTP and literature parts of the CPDB.

Another factor that influences the outcome of carcinogenicity assays is the route of administration. Unfortunately, splitting the CPDB in respect to administration routes resulted in data sets too small for a detailed analysis. Thus, we currently do not know whether or not the route of administration plays a role in the concordance of results.

Compounds that are carcinogenic in rats and mice are considered more harmful than those affecting only one species. Therefore, we grouped the compounds into one-species carcinogens or two-species carcinogens and compared the classifications of the two data sets. For carcinogenic compounds, 48% (16/33) were one-species carcinogens and 52% (17/33) were two-species carcinogens in the NCI/NTP studies; in the literature, 73% (24/33) were one-species carcinogens and 27% (9/33) were two-species carcinogens. A comparison of NCI/NTP data with data from the literature showed that 58% (19/33) of the compounds were classified concordantly (Table 6), but with low coefficients of association (Table 3) and without an indication for a better concordance as obtained for the overall data set. From concordant one-species carcinogens, only 31% (4/13) affected the same species.

*Target-organ specificity.* To facilitate target organ comparisons between rats and mice, we grouped the tissue codes for the target sites into 11 basic target categories: 1, digestive system; 2, liver; 3, cardiovascular system; 4, endocrine system; 5, hematopoietic system; 6, integumentary system; 7, nervous system, brain, and sensory organs; 8, reproductive system; 9, respiratory system; 10, urinary tract; and 11, other (body regions, muscle, skeleton, etc.) excluding some unspecific tissue codes [all tumors, more than one tumor type; tumor types specified in published paper (mix), more than one tumor type; combined by NCI/NTP (MXA), more than one tumor type; combined by Berkeley (MXB) and tumor, or more than one tumor type; and tumor types not specified in paper (*2*)]. For 33 common carcinogens in NCI/NTP and in literature, the liver is the most frequent target site, which is in agreement with other investigations (*2*) (Figure 1).

Subsequently, we classified the compounds either as one-category carcinogens or as multicategory carcinogens, because multisite and multispecies animal carcinogens are considered to pose a greater threat to humans than single-site/species carcinogens (*11*). Our analysis showed that in the NCI/NTP part, the minority (36%; 12/33) of the chemicals was classified as one-category carcinogens and the majority (64%; 21/33) was multicategory carcinogens. In the literature we found similar results: 42% (14/33) of compounds were classified as one-category carcinogens and 56% (19/33) were multicategory carcinogens.

A comparison of NCI/NTP data with data from the literature showed that 52% of the chemicals occurring in both data sets were concordantly classified. The majority of the compounds were multicategory carcinogens (Table 7), but less than 50% of these compounds caused tumors in the same categories. (The statistical analysis summarized in Table 3 reveals a low probability that NCI/NTP and literature results are associated).

*Comparison of species carcinogens and organ-category carcinogens.* Our results indicate a connection between the number of affected species and the number of positive categories. It seems that chemicals classified as one-species carcinogens correlate with one-category carcinogens and that there is also a correlation between two-species carcinogens and multicategory carcinogens.

In both parts of the CPDB, the majority of compounds classified as one-species carcinogens were also one-category carcinogens (NCI/NTP, 9/16; literature, 13/24), and the majority of chemicals classified as two-species carcinogens were also multicategory carcinogens (NCI/NTP, 14/18; literature, 8/9; Table 8), but only 33% of the compounds were grouped concordantly in both parts.

*Quantitative effects.* For the evaluation of carcinogenic potency, we used the most potent TD$_{50}$ for each compound. The TD$_{50}$ is defined as the dose rate in milligrams per kilogram of body weight per day, which if administered chronically for the standard life

**Table 6.** A comparison of one-species and two-species carcinogens in the NCI/NTP and literature parts of the CPDB

| | One species[a] | Two species[b] | Literature |
|---|---|---|---|
| One species[a] | 13 | 11 | 24 |
| Two species[b] | 3 | 6 | 9 |
| NCI/NTP | 16 | 17 | 33 |

Concordant classification: 19 compounds (57%); discordant classification: 14 compounds (43%).
[a]Compounds are carcinogenic in one species (mouse or rat). [b]Compounds are carcinogenic in both species (mouse and rat).

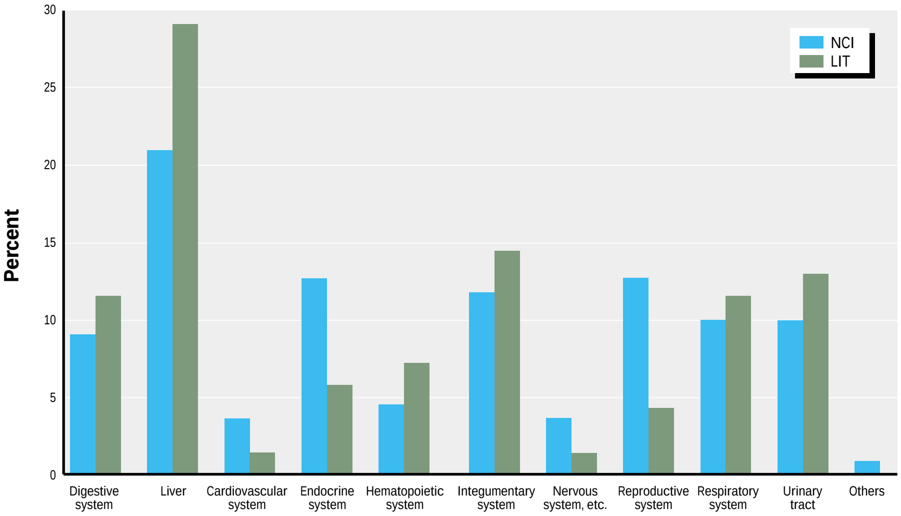

**Figure 1.** Frequency of target organs in the NTP/NCI and the literature part of the CPDB.

span of the species will halve the probability of remaining tumorless throughout that period (*5*). A low $TD_{50}$ value indicates a potent carcinogen, whereas a high value indicates a weak one. If there is only one positive test on the chemical in the species, then the most potent $TD_{50}$ value from that test is reported in the CPDB. When more than one experiment is positive, in order to use all the available data, the reported potency value is the harmonic mean of the most potent $TD_{50}$ values from each positive experiment (*2*).

A comparison of the compounds present in both (NCI/NTP and literature) parts of the CPDB showed that the correlation of quantitative data is similar to qualitative classifications, rather low ($r^2$ = 0.63; Figure 2). This comparison may underestimate the reproducibility of carcinogenicity experiments because, as mentioned several times, only NTP/NCI experiments were conducted with a standardized protocol.

Among chemicals positive in both species (NCI/NTP, 16, literature, 6), the experiments on rats are not only more reproducible but also much more sensitive than tests on mice. These results are in accordance with calculations on the whole CPDB, carried out by Gold and Zeiger (*2*). Furthermore, in the NCI/NTP part, the most potent $TD_{50}$ values were much lower for rats, in contrast to the results in the literature part, where the $TD_{50}$ values for mice were more potent (Table 9).

*Mutagenicity effects.* Mutations are one of the most important mechanisms in chemical carcinogenesis (*4,11*). The *Salmonella* mammalian microsome mutagenicity (Ames) test was designed to measure mutations using several strains of the *Salmonella typhimurium* (*12*). It is well known from the literature that the *Salmonella* assay identifies a high proportion of carcinogenic chemicals, but a number of carcinogens lack mutagenic potential in this assay. Trying to identify the nongenotoxic chemicals with carcinogenic potential represents a major unsolved problem, which shows that rodent carcinogenicity tests cannot, at present, be replaced by the *Salmonella* assay or other short-term tests.

For our comparisons we used the Ames test evaluation from Gold and Zeiger (*2*). The number of compounds occurring in NCI/NTP and literature with additional mutagenicity data was too small for calculations; therefore, we used data from the larger non-overlapping sets. Mutagenicity data were available for 178 chemicals from the NCI/NTP part and for 272 compounds from the literature part. In both subsets the majority of carcinogenic compounds was also mutagenic; 57% (102/178) of the chemicals in the NCI/NTP experiments and 64% (173/272) in the literature investigations were Ames positive. The majority of noncarcinogenic compounds gave negative results in the *Salmonella* assay (119/169 NCI/NTP, 94/149 literature; Table 10).

For further analysis we used only the NCI/NTP data set because only these experiments were conducted under standardized conditions. We tried to determine if mutagenicity data can be used to identify multispecies carcinogens and multicategory carcinogens. The data from the NCI/NTP showed that the majority of one-category carcinogens (52%; 44/85) as well as one-species carcinogens (55%; 54/98) were Ames negative, whereas the majority of multicategory carcinogens (66%; 61/93) and two-species carcinogens (73%; 58/80) were Ames positive (Table 11).

This analysis confirms that genotoxic carcinogens are generally characterized by an ability to cause tumors in multiple species and at multiple sites (*13,14*) whereas nongenotoxic agents tend to exhibit tissue- and species-specific carcinogenicity (Table 11) (*4,13,15*). Results of the *Salmonella* test are therefore important for identifying human carcinogens and to the development of SAR models. The high probability of inducing carcinogenic effects in multiple species is also demonstrated by the majority of chemicals that have been shown to cause human cancers (*11*).

## Discussion

The main observation of our investigation was an unexpected large discordance between experimental results from the NCI/NTP and the literature parts of the CPDB. This leads to two conclusions. First, differences in experimental protocols of the NCI/NTP and literature parts have led to discordant results. NCI/NTP experiments were performed under standardized conditions, but in the literature, experimental protocols vary considerably [although they must meet quality criteria for inclusion in the CPDB (*2*)]. Some chemicals have been tested in the literature part only with one sex of one species, while others have multiple tests that include both sexes and several strains of rats and mice. Additionally, the number of doses, the range of doses, the administration route, and the group size may vary. Thus, different experimental protocols and missing and additional experiments may be the reason for the low concordance between both parts of the CPDB. We were able to demonstrate, however, that the inclusion of more (species, sex, strain, organ) specific information did not improve the reproducibility of the results.

This may be taken as an indication that rodent carcinogenicity assays are, in general, poorly reproducible. This is in clear contradiction to earlier results obtained by Gold et al. (*16*), where an overall reproducibility of 93% (rats) and 76% (mice) was estimated. That investigation was based on replicate experiments of 38 compounds with hamsters, rats, and mice published in the general

**Table 7.** A comparison of one-category and multicategory carcinogens in the NCI/NTP and literature parts of the CPDB.

| | One category[a] | Two categories[b] | Literature |
|---|---|---|---|
| One category[a] | 5 | 9 | 14 |
| Two categories[b] | 7 | 12 | 19 |
| NCI/NTP | 12 | 21 | 33 |

Concordant classification: 17 compounds (52%); discordant classification: 16 compounds (48%).
[a]Compounds are carcinogenic in one tissue category. [b]Compounds are carcinogenic in more than one tissue category.

**Table 8.** A comparison of one- and two-species carcinogens and one- and multicategory carcinogens in the NCI/NTP and literature parts of the CPDB.

| | One/one[a] | One/multi[b] | Two/one[c] | Two/multi[d] | Literature |
|---|---|---|---|---|---|
| One/one[a] | 2 | 4 | 2 | 5 | 13 |
| One/multi[b] | 4 | 3 | 1 | 3 | 11 |
| Two/one[c] | 1 | 0 | 0 | 0 | 1 |
| Two/multi[d] | 2 | 0 | 0 | 6 | 8 |
| NCI/NTP | 9 | 7 | 3 | 14 | 33 |

Concordant classification: 11 compounds (33%); discordant classification: 22 compounds (67%).
[a]Compounds are one-species carcinogens and one-category carcinogens. [b]Compounds are one-species carcinogens and multicategory carcinogens. [c]Compounds are two-species carcinogens and one-category carcinogens. [d]Compounds are two-species carcinogens and multicategory carcinogens.

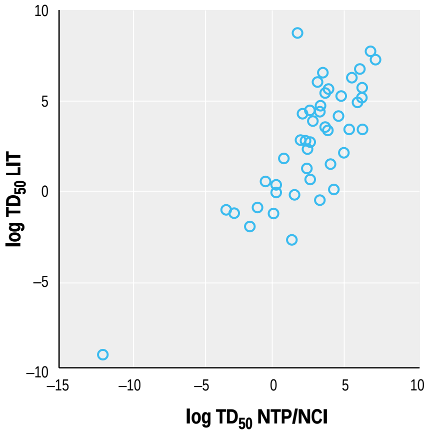

**Figure 2.** Correlation of carcinogenicity $TD_{50}$ values from the NTP/NCI and the literature (LIT) part of the CPDB ($r^2$ = 0.63).

literature. Looking for an explanation for the discordance with our results, we realized that from 47 concordant experiments (sex, administration route, and target organs were considered; therefore, the number of experiments is larger than the number of compounds) with rats and mice listed by Gold et al. (*16*), 34 results were published by the same authors. This may have led to a bias towards identical results, but it may be also an indicator of the importance of strict experimental protocols for reproducibility. In addition, the results may differ for statistical reasons caused by the different data sets (size and selection of compounds).

Based on our data, it is currently impossible to decide between these two explanations, but it seems that the reliability of rodent carcinogenicity assays was overestimated in previous investigations. To clarify this point, it will be necessary to compare a sufficient number of results from standardized assays, which are not publicly available at this time. Until then, results from rodent carcinogenicity assays should be treated as unreliable, which has consequences for SAR modelers and the risk assessment process.

SAR models for carcinogenicity are based on rather poor biologic data. The accuracy of these SARs is therefore much lower than for other end points. It is interesting to note that experiments with rats are more reproducible than those with mice. It is therefore prudent to develop species-specific SAR models. The inclusion of more specific (e.g., target organs) information may make sense from a biologic viewpoint, but this data set is very unreliable and may reduce the overall performance of the SAR model. We assume that data from standardized tests are more reliable than that from nonstandardized sources. When adding data from nonstandardized sources to standardized ones (e.g., NCI/NTP), it should be therefore carefully considered whether the increased amount of data outweighs the additional variability introduced by data from different sources. As mutagenicity is a good indicator, especially for multispecies and multisite carcinogens, this information should be used in SAR studies. An important direction for further research is to find methods to incorporate uncertainty indicators in SAR models and to adequately report the limitations of the derived models.

In risk assessment, the consequences of poor data quality are even more pronounced. We were able to demonstrate that it is not only hard to identify carcinogens in general, but also to identify powerful multispecies and multiorgan carcinogens reliably. Especially in this area, highly relevant to public health and economy, improved quality assurance methods for biologic assays are urgently needed. Furthermore, the uncertainty of biologic information should be adequately considered in the risk assessment process.

**Table 9.** A comparison of the most potent $TD_{50}$ values (medians) for those chemicals classified as carcinogens in the NCI/NTP and literature parts of the CPDB.

| | Rats | | | Mice | | |
|---|---|---|---|---|---|---|
| Data set | All | Males | Females | All | Males | Females |
| NCI/NTP | 6.478 | 2.365 | 7.420 | 32.950 | 54.100 | 29.400 |
| Literature | 50.750 | 161.400 | 70.630 | 20.680 | 28.600 | 15.070 |

**Table 10.** The distribution of mutagens and nonmutagens on carcinogens and noncarcinogens in the NCI/NTP and literature parts of the CPDB.

| Data set | Classification | Mutagens | Nonmutagens |
|---|---|---|---|
| NCI/NTP | Carcinogens | 57% (102/178) | 43% (76/178) |
| NCI/NTP | Noncarcinogens | 30% (50/169) | 70% (119/169) |
| Literature | Carcinogens | 64% (173/272) | 36% (99/272) |
| Literature | Noncarcinogens | 37% (55/149) | 63% (94/149) |

**Table 11.** The distribution of mutagens and nonmutagens on carcinogens and noncarcinogens in the NCI/NTP and literature parts of the CPDB.

| | Mutagenicity | No. of compounds | Percentage |
|---|---|---|---|
| One-species carcinogens[a] | Yes | 44/98 | 45 |
| | No | 54/98 | 55 |
| One-category carcinogens[b] | Yes | 41/85 | 48 |
| | No | 44/85 | 52 |
| Two-species carcinogens[c] | Yes | 58/80 | 73 |
| | No | 22/80 | 27 |
| Multicategory carcinogens[d] | Yes | 61/93 | 66 |
| | No | 32/93 | 34 |

[a] Compounds are carcinogenic in one species (mouse or rat). [b] Compounds are carcinogenic in one tissue category. [c] Compounds are carcinogenic in both species (mouse and rat). [d] Compounds are carcinogenic in more than one tissue category.

Summarizing our experience with data quality in predictive toxicology, we conclude that biologic data are much less reliable than chemical data (*1*). It is impossible to evaluate the quality of standardized rodent carcinogenicity experiments, but results from the general literature have only a poor concordance with results from the NTP. An independent assessment of standardized carcinogenicity assays (e.g., in a round-robin test) is urgently needed to estimate the real reproducibility of these tests. We hope that this paper raises awareness about data quality issues and its implications in predictive toxicology and risk assessment.

## References and Notes

1. Helma C, Kramer S, Pfahringer B, Gottmann E. Data quality in predictive toxicology: identification of chemical structures and calculation of chemical properties. Environ Health Perspect 108:1029–1033 (2000).
2. Gold LS, Zeiger E. Handbook of Carcinogenic Potency and Genotoxicity Databases. Boca Raton, FL:CRC Press, 1997.
3. Available: ftp://helma.informatik.uni-freiburg.de/pub/data/cpdb/ [last update 18 May 2000].
4. Ashby J, Tennant RW. Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. Mutat Res 257:229–306 (1991).
5. Peto R, Pike MC, Bernstein ML, Gold LS, Ames BN. The $TD_{50}$: a proposed general convention for the numerical description of the carcinogenic potency of chemicals in chronic-exposure animal experiments. Environ Health Perspect 58:1–8 (1984).
6. Sawyer S, Peto R, Bernstein L, Pike MC. Calculation of carcinogenic potency from long-term animal carcinogenesis experiments. Biometrics 40:27–40 (1984).
7. Bowen DL, Byrd L, Pereira FCN, Pereira LM, Warren DHD. Sicstus Prolog 3.8.5 User's Manual. Available: http://www.sics.se/isl/sicstus.html [cited 29 March 2001].
8. Holley JW, Guilford JP. A note on the G-Index of agreement. Educ Psychol Meas 24:749–753 (1964).
9. Bortz J, Lienert GA, Boehnke K. Verteilungsfreie Methoden in der Biostatistik. Berlin:Springer, 1990.
10. Evans JS, Gray GM, Sielken RL Jr, Smith AE, Valdez-Flores D, Graham JD. Use of probabilistic expert judgement in uncertainty analysis of carcinogenic potency. Regul Toxicol Pharmacol 20:15–36 (1994).
11. Tennant RW, Zeiger E. Genetic toxicology: current status of methods of carcinogen identification. Environ Health Perspect 100:307–315 (1993).
12. Ames BN, Durston WE, Yamasaki E, Lee FD. Carcinogens are mutagens: a simple test system combining liver homogenates for activation and bacteria for detection. Proc Natl Acad Sci USA 70:2281–2285 (1973).
13. AshbyJ, Tennant RW. Chemical structure, *Salmonella* mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP. Mutat Res 204:17–115 (1988).
14. Ashby J, Tennant RW, Zeiger E, Stasiewicz S. Classification according to chemical structure, mutagenicity to *Salmonella* and level of carcinogenicity of a further 42 chemicals tested for carcinogenicity by the U.S. National Toxicology Program. Mutat Res 223:73–103 (1989).
15. Tennant RW, Stasiewicz S, Spalding JW. Comparison of multiple parameters of rodent carcinogenicity and *in vitro* genetic toxicity. Environ Mutagen 8:205–227 (1986).
16. Gold LS, Wright C, Bernstein L, de Veciana M. Reproducibility of results in near-replicate carcinogenesis bioassays. J Natl Cancer Inst 78:1149–1158 (1987).