5
Extrapolation From Study-Generated Data to Estimates of Human Risk

Contents

LIST OF TABLES

LIST OF FIGURES

5
Extrapolation From Study-Generated Data to Estimates of Human Risk

INTRODUCTION

Information about carcinogenicity is obtained from exposing animals to measured doses of suspect substances in the laboratory or by studying associations between exposures to suspect carcinogens and development of cancer in humans. In practice, both the animals and certain groups of humans, particularly those exposed in the workplace, are exposed to doses far larger than those encountered by most citizens. A number of "numeric extrapolation" methods have been developed to estimate the effect of exposure to low doses based on observed effects at high doses. When information about carcinogenicity is obtained from animals, "biologic extrapolation" techniques are employed to project from animal results to estimates of human risk. (In this report the word "hazard" is applied to a substance or exposure that harbors a "risk" to people who come in contact with it—i.e., a carcinogenic chemical is itself a hazard. Risk is the probability of cancer developing as a result of a particular exposure to the hazard.)

In most cases, no adequate human data are available for estimating risks, and it is necessary to make both numeric and biologic extrapolation from measured responses in animals to estimate human risk. Less uncertainty attends making numeric extrapolations from observation of human responses at high-exposure levels than extrapolations from animal data. However, such extrapolations are often complicated by poor exposure data.

Extrapolation, like testing, is employed to make regulatory decisions, and as in the case of testing, a number of agencies have made statements about the methods they will employ. These policy statements are necessary to explain, in the absence of agreed-on universal procedures, how the agencies intend to use bioassay and epidemiologic data to estimate risk through the use of extrapolation methods.

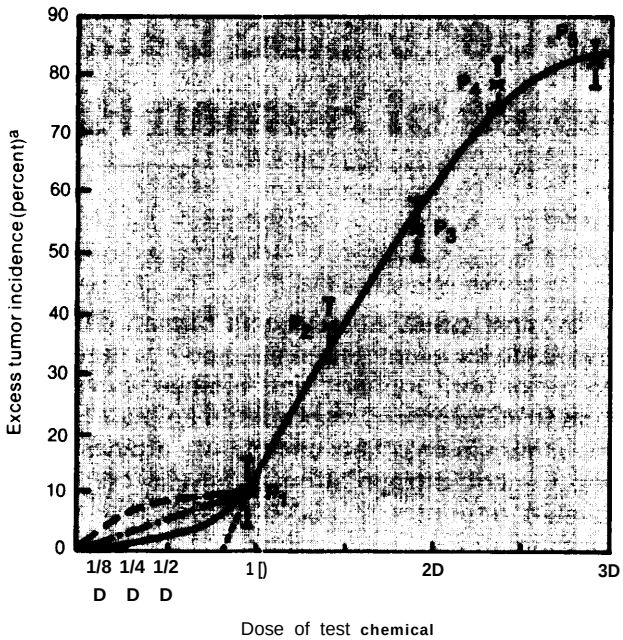
NUMERIC EXTRAPOLATION

This discussion describes information that can be obtained from extrapolation, and comments on various extrapolation methods. It is neither rigorous nor inclusive, and the interested (and mathematically sophisticated) reader is referred to Heel et al. (170), Crump et al. (75) and the Food Safety Council (FSC) (125) for such treatments.

Toxicity testing produces data relating tumor incidence (I) to dosage (D) as shown in figure 21. Generally, a smooth curve drawn between the experimental points, P_1-P_5, (solid line in figure 21) is sigmoidal, or S-shaped. It can be seen that the incidence of tumor formation decreases with decreasing dosage. The crux of the extrapolation problem is what sort of line best approximates the response in the region for which data are not available. Or, what kind of line should be drawn from point P_1 to lower, unmeasured, response levels.

Graphic representations such as figure 21 do not fully show the difficulties in estimating in-

Figure 21.—A Stylized Dose-Response Curve and Some Extrapolated Curves



aExcess tumor incidence (percent) is defined as:

$$\frac{\text{tumors in exposed population}}{\text{number of exposed population}} - \frac{\text{tumors in control population}}{\text{number of exposed population}} \times 100$$

— a sigmoid dose-response curve; infralinear between O and P,
-·-·- linear extrapolation
——— supralinear extrapolation
------ line projected to a threshold

SOURCE: Office of Technology Assessment

cidence at very low doses. The first division on the vertical incidence scale is 10 percent, which means that 1 human or animal out of 10 developed cancer. For many agents, especially those present in air or water, we are interested in knowing what dose is projected to cause an incidence orders of magnitude less, e.g., 1 tumor in 100,000 animals or humans. Such small fractions cannot be seen on the figure, but they can be calculated using **any** extrapolation method.

The solid curved line drawn from point P_1 to the origin is a continuation of the curve constructed between the experimentally determined points. It was drawn by eye, and it is representative of a number of smooth, concave upward lines that can connect P_1 and the origin as a continuation of the sigmoidal curve constructed between P_1 and P_5.

The Question of Thresholds

The solid line on figure 21 embodies the premise that there is no threshold. A threshold model would have the curve hit zero incidence at some dose greater than zero, as is shown in figure **21.**

The threshold argument contends that there are doses of carcinogens so low that they will not cause cancer, and that no matter how many animals are exposed to doses that low or lower, no tumors will result. The counterargument is that any dose of a carcinogen, no matter how small, has a finite although small, chance of causing a tumor, and, if an experiment were performed with a sufficiently large number of animals, such risk would be detectable.

The concept of a population threshold, which is discussed here, includes the idea that there are exposure levels below which no individual in the population will develop cancer. No solution to the threshold/no threshold argument can be found by doing increasingly larger experiments with more and more animals. If a given dose of chemical causes no excess tumors in 1,000 animals, there is no guarantee that it will not cause an excess when 2,000 are exposed.

Individuals may have thresholds, as suggested by the fact that all heavy cigarette smokers do not develop lung cancer. Possible biologic reasons advanced to explain such differences in susceptibilities include physiologic and genetic variations among people. Another reason for the apparent differences in susceptibility may be chance. Some heavy smokers may be “luckier” than others in that their exposures do not trigger carcinogenesis, and they may not develop cancer, or they may die from some other cause before cancer develops. In any case, it is as yet impossible to predict an individual’s threshold for even a single carcinogenic agent and impossible to derive a population’s threshold with current knowledge or methods.

In a highly recommended article, Maugh (224) reports on conversations with a number of scientists and administrators from the National

Institutes of Health, private testing laboratories, and industry. The conclusion of the article, apparently shared by the interviewed scientists and administrators, is that resolution of the threshold question is not now available:

> . . . it is extremely unlikely that it would be possible to distinguish between a linear dose-response curve and a highly nonlinear one (threshold), even in a large-scale experiment involving several thousand animals per dose level.
>
> . . . statistical analysis of standard animal carcinogenicity experiments, Schneiderman [then-Associate Director of the National Cancer Institute (NCI)] concludes, does not now, and probably never will, resolve the threshold question, There are, he says, simply too many "biologically reasonable" mathematical models, both implying and denying the existence of thresholds, that will fit the observed results.
>
> . . . there is so little data and so many interpretations, Gehring [Dow Chemical Co.] says, arguing about thresholds is an exercise in futility.

The view that thresholds cannot be demonstrated is accepted in publications of the **National** Research Council (NRC) (262), Interagency Regulatory Liaison Group (IRLG) (180), FSC (125), and, for tumors resulting from somatic mutations, by the American Industrial Health Council (AIHC) (8). Despite the apparent agreement that it is impossible to demonstrate a threshold, many individuals object to the idea that thresholds should not be considered in making decisions about carcinogens.

In particular, some tumors are thought to result from irritation or mechanical injury, and threshold models are postulated for those. An often mentioned example is Clayson's (60) finding that bladder stones are found in conjunction with some bladder tumors. Stone formation may be dependent on intake of large quantities of a chemical; if exposure to the chemical is sufficiently low so that no stones are formed and if stone formation is necessary for carcinogenicity, then a threshold should exist. This and other examples of "epigenetic" tumors can be considered separately from tumors that originate from somatic mutations. Just as thresholds for acute toxic responses differ, thresholds for stone formation most likely differ among individual animals. Therefore, it may be impossible to determine exactly what dose is necessary to produce a stone and therefore pave the way to a tumor. Nevertheless, if the threshold for stone formation in animals is found to be 100 or 1,000 times greater than human exposure levels, the observation takes on great significance. If stone formation is a necessary precursor to tumor appearance, and no stone formation occurs at human exposure levels, it becomes difficult to maintain the position that human exposure levels present a carcinogenic risk. The difficulties in this argument are the possibility that some humans may produce stones at very much lower doses than animals and that carcinogenicity might not proceed through stone formation in humans. Such caveats seem reasonable to some people; unreasonable to others. Experimentation and more sophisticated models may eventually settle such questions. However, for the present, positions on these issues reflect organizational policy and individual judgment.

Gehring, Watanabe, and Young (139) describe differences between metabolism of chemicals administered at low and high doses. Measurement of these differences is called pharmacokinetics. In general, biochemical mechanisms to detoxify chemicals and to repair damage to DNA are seen as having a better chance of detoxifying and repairing damage at low doses than at high doses. High doses can swamp defense mechanisms resulting in toxic effects; lower doses are seen as presenting little or no risk.

Metabolic differences measured by pharmacokinetics might have important effects on the shape of the dose-response curve. If a metabolize of an ingested substance is the actual carcinogen, and the production of the metabolize increases out of proportion to dose above a certain level, then the dose-response curve might bend upward at that point. Pharmacokinetic data are not always collected, and the standard bioassay which produces data at only two dose levels does not provide sufficient information to see if differences in metabolism might be important in the slope of the dose-response curve.

Biochemistry of a particular compound may be affected by other compounds present in the organism. Cancer is a relatively common disease, and if avoiding cancer involves biochemical detoxification processes, it may be that small doses of many substances can swamp the process as well as large doses of a single substance. If this is correct, the addition of the carcinogens, even at low doses, might be enough to overcome the detoxification mechanism.

Cornfield (67) described quantitative extrapolation models to take into account differences in metabolism at high and low doses. His paper was criticized because it postulated a threshold (see 36,73,229,276,.319), and his hope that the paper would lead to discussion of the merit of his models apparently was not realized (68). The difficulty with models which propose detoxification activities for producing thresholds is that the detoxification has to be instantaneous and complete. Otherwise molecules might escape detoxification and initiate a carcinogenic event.

Donald Kennedy, a few months after leaving the post of Commissioner of the Food and Drug Administration (FDA), expressed his opinion that thresholds may exist for some chemicals:

> Dr. Kennedy said that the [Delaney] clause in the FDA's authorizing legislation codifies the hypothesis that there is no threshold concentration below which a chemical does not cause cancer. And although this hypothesis "probably holds most of the time," Dr. Kennedy said that he was "as certain as I can be of any scientific prediction that some day, very soon, some compound will be demonstrated to have a threshold level for cancer causation" . . . (274).

It is difficult, if not impossible, to marshal more evidence on one side of the threshold question than on the other. The ascendancy of the more conservative view, that thresholds cannot be identified for human populations, can be taken as a policy decision made in the interest of protecting the public health. Such a general policy can not exclude the possibility that a threshold may someday be demonstrated.

Numeric Extrapolation To Project Risk at Doses Below Those Tested

The shape of the line in figure 21 depends on the number of tumors observed at points P_1-P_5. No matter what method is used to extend the line below P_1, that extension represents an estimate. Any number of smooth curves can be drawn from point P_1 to the origin; for convenience, the possible lines will be divided into three families: supralinear, linear, and infralinear. A detailed discussion of these models as they relate to radiation and cancer is available in a paper by Sinclair (329).

Supralinear Extrapolation

A supralinear extrapolation is presented on figure 21. It says that some doses less than ID are relatively more effective in inducing tumors than doses equal to ID. Conceptually the contention that lower doses are more carcinogenic is easy to address. Further tests at lower doses would resolve the question, but additional tests are costly and time consuming.

Supralinear models are considered for two reasons. In several NCI bioassays, the tumor yield was lower at the high dose than at the low dose (102). In other words, the lower dose was more efficient at producing tumors. The explanation is that the higher dose was so toxic that it killed animals before they developed tumors. The other reason for considering supralinear responses is that some studies of radiation-induced cancer have been interpreted as producing supralinear dose-response curves (e.g., 241), but those interpretations are hotly disputed.

Such responses might result from the presence of a subpopulation of more sensitive individuals. On figure 21, the supralinear response between the origin and P_1 represents the tumors induced in the proposed sensitive fraction of the population; the solid line drawn from the origin to P_5 represents the sensitivity of the remaining members of the population. The difference between the two lines between the origin and P_1

represents the contribution of the sensitive subpopulation to the total response at doses below ID. It can be seen that the sensitive subpopulation accounts for the majority of tumors that *occur* below P_1. Clearly, if this model describes risks, reducing doses to 1/2D or 1/4 D would not significantly reduce tumor incidence, and a supralinear dose-response curve would force lowering doses to very small fractions of D to significantly reduce tumor incidence.

Nonartifactual, supralinear dose responses have rarely been observed in bioassays but neither would they be expected. Laboratory animals are highly inbred and each animal should be more nearly equally sensitive than are members of human populations. Supralinear response models have been advanced but do not now receive the acceptance accorded to the other two general models.

Linear Extrapolation

A linear model is shown by the straight line that extends from P_1 to the origin on figure 21. If the true dose-response curve is represented by the solid curved line from P_1 to the origin, then the linear model is "conservative" and overestimates the number of tumors at all doses between P_1 and the origin.

The paper by Crump et al. (75) is an oftencited and important argument for linear dose responses at low doses. The paper points out that **25** percent of the U.S. population will develop cancer as a result of existing carcinogenic influences (see ch.3). Crump et al. (75) propose that any new carcinogenic substance interacts additively with exposures and behaviors already present in the environment. Their mathematical theories predict that regardless of the shape of the dose-response curve at high exposures, at low doses cancer incidence should be proportional (linear) with exposure to the substance under study.

Gaylor and Kodell (137) argue that no risk estimate can be very reliable for doses below that associated with the lowest data point (P_1 in figure 21) because there is no information available below point P_1. They propose the use of "linear interpolation" and the 95-percent upper confidence level to estimate the *maximum* risk posed by a substance.

Error is associated with any experimental determination, and standard methods can be used to calculate "confidence limits" for each estimate. Usually "95-percent confidence limits" are calculated for carcinogenicity experiments; they are plotted as vertical bars extending from the data points, as shown on the figure. The **95-**percent confidence limit says that given the experimentally determined incidence and the size of the experiment, we can be 95-percent certain that the actual incidence represented by the point estimate lies inside the error limits.

In the method of Gaylor and KodelI (137) a line is drawn from the upper limit of the error bar on the lowest data point to the origin. Inspection of figure 21 shows that this method projects a larger risk than does linear interpolation from the point P_1 to the origin. This is not an estimate of risk; it is an estimate of the upper bound of risk.

Objections to including upper confidence levels in extrapolation are frequently voiced. The practice of including them is seen as introducing a "safety factor." Industry spokesmen (9a) and others contend that the best risk estimate should be made and the safety factors added after the estimate is made.

As a practical matter, there is often no alternative to the linear model. The dose-response curve in figure 21 is an outrageous overstatement of the data that are generally available. Bioassays carried out according to NCI's cancer testing guidelines (331) produce only two data points. The Environmental Protection Agency's (EPA) (102) analysis of many such tests showed that tumor incidence was sometimes higher and sometimes only measurable at the lower of the two doses because other toxic effects killed animals at the higher dose. Left with only the response at the lower dose, there is little choice available but to estimate responses at still lower doses on the basis of simple proportionality. Such calculations produce a straight line from the experimental point to the origin.

IRLG (180) did not discuss how to extrapolate when more than one data point is available; it

recommends linear extrapolation (proportionality) for making estimates from a single point. IRLG further proposes that the upper confidence level be used as the starting point for extrapolation to achieve "an added degree of protection . . . ."

A linear extrapolation model from the lowest positive data point to zero dose (104) was used by EPA's Carcinogen Assessment Group (CAG) until the summer of 1980. At that time, CAG (48) announced it was going to discontinue use of the linear model and subsequently employ a model developed by Crump. The CAG decision was not made because evidence had shown the linear model was poorer than the new one (48):

> There is no really solid scientific basis for any mathematical extrapolation model which relates carcinogen exposure to cancer risks at the extremely low level of concentration that must be dealt with in evaluating the environmental hazards.

The now-adopted model is linear at low doses, and, in practice, produces estimates of risk at low doses which " . . . are not markedly different from those obtained with the former procedure based on the one-hit [linear] model" (230). However, the new model does allow consideration of data produced above the linear part of the curve (points P_4 and P_5 on figure 21) to influence the slope of the line and the range of error associated with each point.

Infralinear Extrapolation

The curved line between P_1 and the origin on figure 21 or any curved line which remains below the straight line is infralinear. Such models predict lower tumor incidence than the linear model. If it were decided that a certain level of risk were acceptable, higher exposure to the chemical would be allowable under infralinear than under linear models.

A number of such models have been developed and are well described in FSC's (125) report. All of them produce concave upward lines between the origin and the lowest data point (P_1, in figure 21). Different models produce differently shaped lines. It is suggested (e.g., 125) that the model which produces the line that best fits the data points (P_1 to P_5 in figure 21) is the model to use to predict risk between P_1 and the origin. Unfortunately, usually any of the models seems to fit the available data points about equally well (9a,125,138,180). The reasons for the equally good fits are that generally only one, two, or three data points are available and all of them measure incidence above 10 percent. Data points at such relatively high response rates do not often provide enough information to decide what the dose-response curve is at an incidence of 1 percent or less.

QUANTITATIVE EFFECTS OF SELECTING A MODEL

Selection of the appropriate model for estimating risks at low doses would be made easier if some models clearly did not fit the observed **data points. As mentioned above, hardly ever is it possible to select the best model or even to reject the worst on the basis of fit to observed data points. The** low end of the dose-response curve is most informative for selecting the correct model but it is the part that is most difficult to measure. In practice, incidence rates in animal tests much below 113 percent (5 tumor-bearing animals in a test population of **50)** can seldom be distinguished from the rate of spontaneous tumors.

Table 32, derived from a paper by Brown (35), shows that two infralinear models and the one-hit model, which is essentially linear at doses that cause an incidence of 10 percent or less, are indistinguishable at high doses. For the table, a dose level of one was set as sufficient to cause an incidence of 50 percent. The expected incidence using higher doses or doses as low as one-sixteenth are nearly equal regardless of the

Table 32.—Expected Incidence of Tumors Calculated by Three Models When a Dose of 1.0 Caused Tumors in 50 Percent of the Tested Animals

	Projected percentage of tumor bearing animals		
Dose level	Log-normal model (Infralinear)	Log-logistic model (Infralinear)	Single-hit model (linear at incidence below 10°/0)
16	98	96	100
4	84	84	94
1	50	50	50
1/4	16	16	16
1/16	2	4	4
1/100	0.05	0.4	0.7
1/1 ,000	0.00035	0.026	0.07
1/10.000	0.0000001	0.0016	0.007

SOURCE: Adapted from Brown (35).

model. Brown (35) points out that no experiment of practical size could distinguish among the three models at those dose levels.

However, at much lower dose levels of 1/100, 1/1,000, and 1/10,000, the models diverge greatly in their projections of incidence. These greatly lower dose and response levels are often the ones of most interest for estimating human risks, but they cannot be measured. The incidence measured at higher doses do not provide sufficient information to choose the appropriate model, These problems plague all extrapolation efforts.

In general, either a linear or infralinear model is used for extrapolation. The linear model predicts a higher incidence at low doses than does the infralinear model.

Selection of the correct extrapolation model is important for *only one* of the three possible regulatory strategies for carcinogens. The first strategy is to accept either human or animal evidence as sufficient to identify carcinogens, and once the identification is made, try to eliminate the exposure. This approach requires no quantitative or numeric extrapolation. The second approach uses biologic and numeric extrapolation to rank substances in order from that expected to be most carcinogenic to those that are noncarcinogenic. This relative ranking can be accomplished by consistently applying any model, and the numerical accuracy of the estimated incidence is not critical. The third approach, which includes a quantitative estimate of human risk to be used in risk-benefit computations or to consider levels of acceptable risk requires the most accurate numerical estimate. Clearly, in this case, the selection of models is important because the numbers produced by different models vary across a wide range,

Virtually Safe Doses

A very low risk of cancer, say, one chance in million lifetimes, is sometimes suggested as a virtually safe dose. Any extrapolation model can be used to calculate the dose which will produce such a risk, and different models produce very different estimates for the virtually safe dose (see table 32 and 125,170). As shown on figure 21, infralinear models predict higher virtually safe doses (i.e., lower risks at any dose) than does the linear model.

WHAT QUANTITATIVE PROJECTIONS CAN BE MADE FROM NEGATIVE RESULTS (ZERO EXCESS TUMORS IN A TEST POPULATION)?

No tumors occurring in a test population of 100 animals or no excess tumors among **100** animals as compared with the number of tumors in 100 control animals does not show that zero cancer risk is associated with the chemical. Instead standard statistical calculations based on zero excess tumors in 100 animals show that we can be 95-percent confident that the actual incidence of tumors is no more than 4.5 percent. This estimate of the incidence of tumors that might have gone undetected is called the upper confidence limit. The percentage can be reduced by testing more animals, for instance, finding zero excess tumors in 1,000 animals would mean that we can be 95-percent confident that the actual incidence is no more than **0.45** percent.

Proceeding with the illustration of zero excess tumors in 100 animals, assume that the dose administered to the animals was 1,000 times higher than that to which humans are exposed. Linear extrapolation (proportionality) predicts that exposure of the U.S. population to that chemical at 1/1,000 the level fed to) the animals will result in fewer than 10,000 **cases** of cancer assuming equal sensitivity between man and animals. The estimated risk would be reduced if the experiment on which it is based is more sensitive. For instance, finding zero excess tumors in 1,000 animals (instead of 100) would reduce the estimated risk to fewer than 1,000 cases.

The statistics of the above exercise are not questioned, but they are seldom applied. Although a test cannot show that a substance presents no risk, much less concern is attached to substances that cause no excess of tumors. The risk associated with substances that are negative in bioassays is qualitatively lower, and the consideration of quantitative risk estimates from negative experiments is of minor importance.

OTHER EXTRAPOLATION MODELS

The supralinear, linear, and infralinear models are all dichotomous. They compare the number of tumors or tumor-bearing animals in the exposed population to the number in the controls. In both populations, the analysis depends on the presence or absence of tumors. Other models can be used to make inferences about the times (or ages) at which animals develop tumors in response to exposures.

Two of these models have been used extensively to describe animal and human "time-to-tumor" data. The lognormal model described in Chand and Hoe] (57), predicts that the average time-to-tumor is longer at low doses. An important outcome of this model is that at sufficiently low doses, the time necessary for tumor development may exceed the expected lifespan. Such a long latent period would produce a "practical threshold." The Weibull model (also described in 57) predicts that the average time-to-tumor is nearly independent of dose. This prediction means that an increase in dose simply causes more cancers; it does not shift the age distribution at which they occur. The assumptions and predictions of these two models are quite different. Unfortunately both are apt to give adequate fits to any available data set, making it difficult to reject one in favor of the other. IRLG (181) concluded that these models have not received the attention that has been focused on dichotomous dose- incidence relation models, and it recommends more research be directed toward exploring them. The mathematics for these models is sophisticated, and the interested reader is referred to Chand and Heel (57).

At this time, discussion of other models is more an academic than a policy exercise. Opposing camps **are** for or against quantitative ex-

trapolation, and among those favoring it, the argument is between those for and against using linear extrapolation (including EPA's new model). There is no agreement about another single model being offered as an alternative at this time.

EXTRAPOLATION FROM SHORT-TERM TESTS TO HUMAN RISKS

McCann et al. (227) and McCann and Ames (226) showed that about 90 percent of the carcinogens tested in the Ames test were mutagenic (see ch. 4). Meselson and Russell (235) developed a model to compare the mutagenic and carcinogenic potency of tested chemicals. Fourteen chemicals were analyzed because there were sufficient mutagenic and carcinogenic data available to construct dose-response curves for each. The correlation between animal carcinogenicity and bacterial mutagenicity was excellent for 10 of the 14 compounds. The other four compounds (all nitroso-compounds) were more potent as carcinogens than as mutagens.

A spirited exchange of views resulted from the suggestion that quantitative relationships exist between mutagenicity in the Ames test and carcinogenicity in animal tests. Ashby and Styles (18) challenged the idea that such relationships were common, and Ames and Hooper (12) responded that they were.

The International Program for the Evaluation of Short-Term Tests for Carcinogenicity (188) distributed 42 chemicals to each of 12 laboratories for testing in the Ames system: To eliminate bias, none of the laboratories knew the identity of the chemicals. There was excellent agreement among test results obtained in different laboratories and about 80 percent of the carcinogens were scored as mutagens and about 80 percent of the noncarcinogens were scored as nonmutagens. These numbers compare well with those in the literature that describe results from experiments in which the investigators knew before the mutagenicity test was run that the chemicals were or were not carcinogenic. Although the 80-percent is lower than the 90-percent accuracy sometimes reported, the program included a number of chemicals that are known to present difficulties for the Ames test. In a qualitative sense, the test performed very well, but mutagenic potency did not correlate with carcinogenic potency. In other words, the results from this program do not support the idea that there is a quantitative relationship between Ames test results and carcinogenicity in animals. Similar results showing good qualitative and poor quantitative agreements between mutagenicity and carcinogenicity were reported by Bartsch et al. (22).

Meselson and Russell **(235)** reported that sufficient quantitative data were available for two human carcinogens, aflatoxin B and cigarette smoke, to allow comparison of mutagenic potency to carcinogenic potency in humans. Correlation between carcinogenicity in humans and mutagenicity was good for the two compounds, and the Meselson and Russell paper raised the possibility that human cancer risks might be predicted from mutagenicity data. Acceptance of such a procedure is far away and will depend on much more data being available to support the proposed quantitative relationships.

Clearly, controversies now exist about the value of extrapolations made from short-term tests. It seems reasonable that, as use of short-term tests increases, such projections are going to be made, but some initial optimism about the value of quantitative extrapolation from short-term tests to carcinogenicity is apparently fading.

CARCINOGENIC ACTIVITY INDICATORS

The NRC Pesticide Committee **(267)** recommends calculating potency expressions, "Carcinogenic Activity Indicators" (CAI), for tested chemicals. For each point of a dose response curve (fig. 21), the number of chemical molecules ingested divided by the animal body weight can be related to the excess percentage of tumor-bearing animals in the exposed population.

$$\text{CAI} = \frac{\text{excess percentage of subjects in which tumors are observed}}{\text{lifetime dose (molecules/kg of body weight)}} = \frac{\text{tumor incidence}}{\text{dose}}$$

CAIs do not have to be based on total tumors, for instance, site specific tumors may be counted, the analysis may be limited to one sex, only malignant tumors may be counted, or other alterations can be made as wanted. In practice, the number of molecules of different substances, S1, S2, and S3 required to induce the same percentage of excess tumors can be compared to determine which is the most potent carcinogen. CAIs will probably be different for each point on a dose--response curve because the points seldom fall in a straight line, but comparisons can be made at comparable doses.

Using information about exposure levels for human populations, the number of molecules that compare to human exposures can be calculated. Linear extrapolation is then to be used to estimate the animal response at human exposure level. This method is especially appropriate for comparing chemicals with similar uses, and if applied, would assure that the most and least risky ones based on animal data are identified.

The procedure does not make predictions of human cancer risks from animal data and avoids the problems associated with biologic extrapolation. The NRC committee (267) urges that only epidemiologic data be used to estimate human risk; it restricts the use of animal data to making comparisons of carcinogenicity in animals. (Such an approach is especially attractive when deciding about regulating pesticides. Substitutes are often available and CAIs offer a method to decide on the less or least risky one. This suggestion corresponds to the second possible use of extrapolation discussed in *Quantitative Effects of Selecting a Model* above.)

POTENCY

Ames et al. (14) have analyzed more than l,500 bioassays carried out on some 600 chemicals. For each experiment, they have calculated a potency index, TD_{50}, which is calculated as the total dose of substance necessary to produce tumors in **50** percent of the animals. They expect to compare potency:

1. among multiple tests run on the same substance in the same strain and species;
2. between male and female animals;
3. between different strains of the same animal;
4. between different sites in different animals; and
5. between rodent tests and 26 tests that have been carried out: in monkeys.

The results of this massive project is expected to provide much information about biologic extrapolation from species to species and about potencies.

Ames et al. (14) mention that preliminary results show that there is usually less than a tenfold variation in potencies among rodents. This level of agreement was also found by Crouch and Wilson **(72)** for most of 70 chemicals tested in both rats and mice by NCI. Crouch and Wilson included tumors which were not present at "statistically significant" levels in their analyses. Therefore, some chemicals judged positive in only one species by NCI (146) have potencies that agree within a factor of 10 between rats and mice although one is statistically significant and

the other is not. Results showing a substance is much more potent in one species than in another will suggest that metabolism of the chemical differs from species to species. This information will also be important in efforts to improve extrapolation methods.

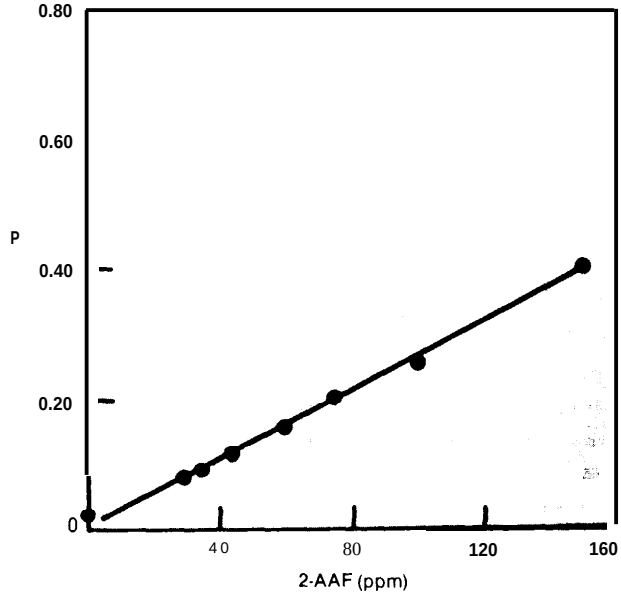
THE ED_{01} EXPERIMENT AND EXTRAPOLATION MODELS

The National Center for Toxicological Research, a joint FDA/EPA laboratory, tested the known bladder and liver carcinogen 2-acetylaminofluorene (2-AAF) in approximately **24,000** female mice. This experiment was designed to study dose responses down to a 1-percent tumor incidence, i.e., the effective dose for a l-percent (0.01) response (ED_{01}).

Figures 22 and 23 show the results of postmortem examination for liver and bladder tumors in animals exposed to 2-AAF for between **21.5** and 28.5 months. The curve that describes the liver tumors (fig. 22) shows no threshold and increases almost proportionally to dose. The curve that describes the bladder tumors (fig. 23) gives an impression of a "threshold" dose below 60 ppm of 2-AAF. Gaylor (136) ascribes the apparent threshold to a lack of resolution of the graph for low tumor rates. He replotted these same data on an enlarged scale (figure **24)** to show that the number of bladder tumors increases with dose even at lower doses. This diagram supports the idea that there was no threshold. However, there is a dramatic change in slope of the line between 60 and 75 ppm, and the efficiency of 2-AAF in causing tumors increases greatly at doses above 60 ppm (fig. 24).

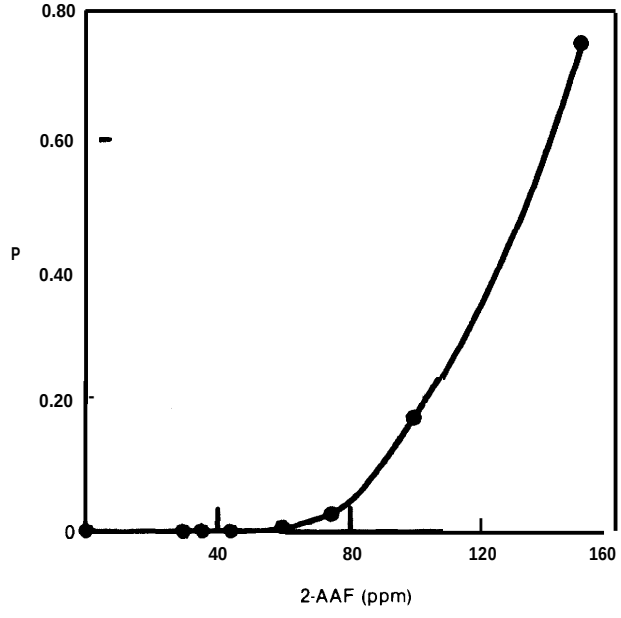
Carlborg (50) has also analyzed the data from ED_{01}. He does not argue that the bladder cancer data show a threshold, but he does contend that neither the bladder nor the liver data fit a one-hit (linear) model. In addition to trying to fit a one-hit model to the data, he also tried the

Figure 22.—Proportion (P) of Mice With Liver Tumors v. Dose (21.5 to 28.5 months)



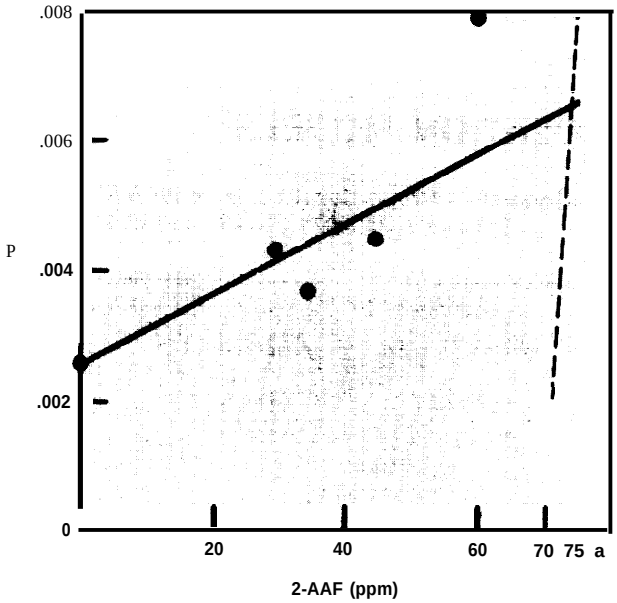
SOURCE Littlefield, et al (213)

Figure 23.— Proportion (P) of Mice With Bladder Tumors v. Dose (21.5 to 28.5 months)



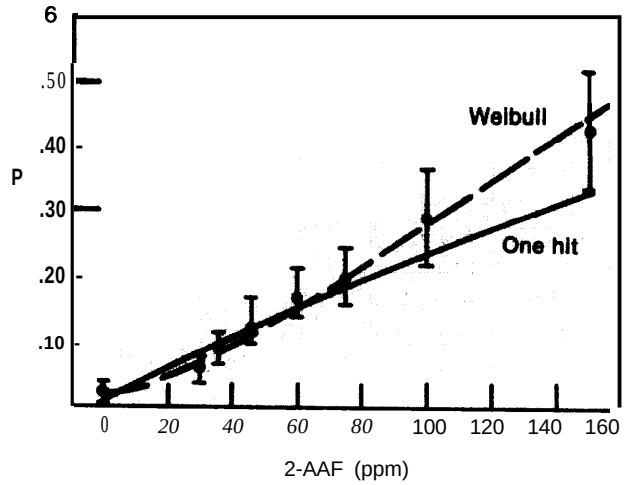
SOURCE Littlefield, et al (213).

Figure 24.— Proportion (P) of Mice With Bladder Tumors v. Dose (21.5 to 28.5 months)



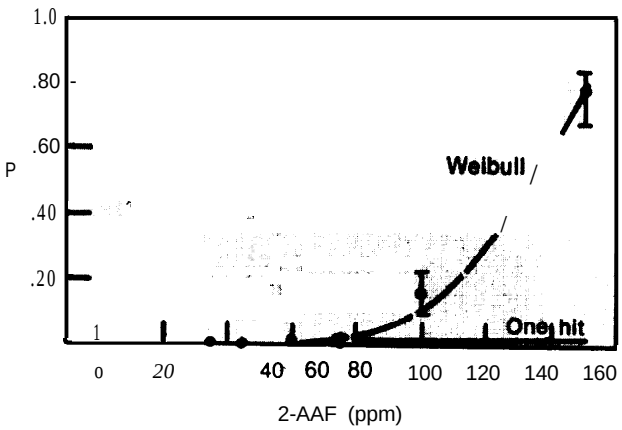
NOTE: The vertical scale has been expanded 100 times compared to figure 23
—— line fitted to data points between 0 and 60 ppm
–– slope of line between 60 and 75 ppm (derived from figure 23)
SOURCE: Gaylor (136) and Office cf Technology Assessment.

Figure 25.—Proportion (P) of Mice With Liver Tumors v. Dose (24 months)



SOURCE: Carlborg (50)

Figure 26.— Proportion (P) of Mice With Bladder Tumors v. Dose (24 months)



SOURCE: Carlborg (50)

Weibull model, a "generalization of the one-hit model." Figures 25 and 26 represent his analysis of the two sets of data, and the "best-fitting one-hit and Weibull models" are plotted in both figures. It can be seen that the Weibull model provides a better fit, especially with the bladder tumor data. Carlborg (50) also states that the one-hit model fits well with only the three lowest dose points from the liver tumor data.

The Weibull model produces a different slope for the line to be drawn from 30 to O ppm. Carlborg calculates that a 0.000045 ppm dose of 2-AAF is necessary to produce a one in a million risk of liver cancer using the one-hit model. Using the Weibull model, the calculated dose for a one in a million risk is 100 times higher, 0.0045 ppm. In other words, use of the Weibull model would allow exposures 100 times higher than the linear model if there were agreement that a one in a million risk were acceptable.

Despite the large number of data points from ED_{01}, Gaylor (136) argues that a "best" model cannot be chosen on the basis of fit with the data and that the Weibull model cannot be singled out above all others. Furthermore, he says that several models fit the observed data about equally well, and that they predict very different levels of risk at doses between O and 30 ppm 2-AAF (see discussion of different models in 35,49,51,123,170).

Obviously, animal data can be best used quantitatively to estimate potential risks within the same animal species under test conditions. These estimates cannot be used directly to estimate human risks. Other "biologic" assumptions are necessary to extrapolate from animal data to estimates of human risk.

BIOLOGIC EXTRAPOLATION FROM ANIMAL TESTS

Bioassay guidelines call for testing in two species. Qualitative judgments are based on whether a tested substance causes tumors in both, only one, or neither species. Sometimes, a substance is reported positive (causes tumors) in only one species and negative in another. The frequency with which this problem is encountered can be estimated from Griesemer and Cueto's (146) discussion of the results of the NCI bioassay program (see app. A). Of the 98 chemicals for which "very strong" or "sufficient" evidence of carcinogenicity was found in either rats, mice, or both, 54 were positive in only one species. An analysis of test results from **250** substances (300) found 38 percent were carcinogenic in neither rats nor mice, 15 percent were carcinogenic in either the rat or the mouse but not both, and 44 percent were carcinogenic in both. Crouch and Wilson (72) and Ames et al. (14) tend to dismiss these discordant results between rats and mice because tests analyzed by their methods, which consider experimental error and test sensitivity, are in agreement much more often.

In cases in which there are apparent species differences in sensitivity, the positive result is generally accepted as more important. The Office of Science and Technology Policy (281), IRLG (180), the Occupational Safety and Health Administration (279), and American Federation of Labor/Congress of Industrial Organizations (7) all present arguments for following this course. AIHC (8) and Purchase (300) present arguments against deciding that the more sensitive animal is predictive for human response. The disagreement continues, but in Federal programs, extrapolation is based on results from the more sensitive species.

Accepting that experimental animals provide appropriate data for extrapolation to estimates of human risk, a decision has to be made about how to adjust the dose measured in the bioassay to the dose experienced by humans. A mouse or rat, of course, is much smaller than a human, and the dose necessary to cause a carcinogenic response is less than that required in humans. Three "scaling factors" are in general use to make allowance for the different sizes and rates of metabolism between experimental animals and humans. The three are listed below in order from least conservative, that is the one that predicts the lowest human risk, to the most conservative. The fourth scaling factor is less often used, but it is included because it was used for the estimates reported in table 34. For rats and mice, the most commonly used laboratory animals, the relationship between scaling factors and estimated risk in man for the same doses are shown in table 33.

1. Exposures may be adjusted on the basis of relative body weights, milligram of agent/kilogram of body weight/day (mg/kg/day), for animals and humans. This method is most generally used by toxicologists.
2. In cases where the experimental dose is measured as parts per million in food, air, or water and human exposure is through ingestion, the dose of the chemical is expressed as parts per million. This method is generally used by FDA and in some cases by EPA.
3. Exposure may be adjusted on the basis of the relative surface areas of the test animal and humans: milligram of agent/surface area/day, for animals and humans. It is generally expressed as milligram of substance/square meter of surface area/day or $mg/m^2/day$. EPA *uses* this scaling factor.

80-480 O - 81 - 12

Table 33.—Relative Human Risk Depending on How Dose Rate is Scaled From Experimental Animals to Humans

	Risk projected for humans when an identical dose is scaled by different factors			
	Milligram	Parts per million	Milligram	Milligram
Experimental animal	Kilogram body weight/day	In diet	m^2 body area	Kilogram body weight/lifetime
Mouse	1	6	14	**40**
Rat	1	**3**	**6**	**35**

SOURCE: Off Ice of Technology Assessment.

4. Exposures may be adjusted on the basis of relative body weight over lifetime, milligrams of agent / kilogram of body weight / lifetime (mg / kg / lifetime).

As can be seen from table 33, the choice of scaling factor can make a difference of up to fortyfold in estimating human risks. The mg / kg / day scaling factor was arbitrarily set equal to 1.0. Use of the mg / m^2 / day factor (for instance) projects that humans would have 14 times the risk of a mouse for equivalent doses measured in mg / kg / day. The information given in table 33 allows a comparison to be made among the scaling factors. However, it is important to remember that great uncertainties surround biologic extrapolation because of possible differences between laboratory animals and man, and no great assurance is attached to any number in table 33.

COMPARING MEASURED HUMAN CANCER INCIDENCE AND MORTALITY TO ESTIMATES MADE USING EXTRAPOLATION

The more troubling and more fundamental problem with biologic extrapolation concerns questions about how closely the test animal resembles humans. This problem is partially related to differences in the greater genetic complexity of human populations. Populations of test animals are highly inbred and are almost genetically identical. Populations of humans are outbred and include greatly differing genotypes. There is no way to deal with the problem of humans that may differ in sensitivities because there is seldom, if ever, a way to associate sensitivities with individuals. The other problem concerns differences in metabolism between test animals and humans. Few are well understood, and many may be unidentified.

Laying these problems aside, a few efforts have been made to compare human cancer incidence or mortality to the levels of incidence or mortality extrapolated from animal studies. The number of such attempts is limited by the few cases for which data are available both from experimental animals and from humans. The Consultative Panel on Health Hazards of Chemical Pesticides of the National Research Council's Study of Pest Control **(262),** identified six chemicals for which such comparisons could be made. Comparisons were made on the basis of lifetime dosage (expressed as milligram chemical / kilogram body weight / lifetime) in animals and in humans. Table 34 shows those findings. It can be seen that for three chemicals the incidence of human tumors was essentially that predicted from animal studies, and for the other three, extrapolation from animal data overestimated human risk as measured by epidemiology. Crouch and Wilson (72) have made similar comparisons for 13 chemicals (including the 6 in table 34). They reported good agreement between predicted and observed human tumors rates, using a linear nonthreshold extrapolation model. Crouch and Wilson **(72)** also calculated "scaling factors" from their experiments and concluded that humans are twice as sensitive as mice and between one-third and three times as sensitive as rats to the same dose expressed as mg / kg / day. These values differ

Table 34.—Comparison of Tumor Rates in Laboratory Test Animals and Humans Following Lifetime Exposures to Comparable Amounts of Each of Six Agents (comparison based on mg agent/kg body weight /lifetime)

Chemical	Test animal	Animal tumor site(s)	Human tumor site	Relative tumor rate[a]
Benzidine	Mouse	Liver	Bladder	ca. 1
	Rat	Bladder		
Cigarette smoking	Mouse	Lung	Lung	ca. 1
	Hamster	Larynx		
N, N-bis(2-chloro-ethyl) -2-naphtyl-amine	Mouse	Lung	Bladder	ca. 1
Aflatoxin B,	Mouse	Liver	Liver	ca. 10
	Rat	Liver		
Diethylstilbestrol (DES)	Mouse	Mammary	Daughters' re-productive tract	ca. 50
	Mouse	Cervix and vagina		
Vinyl chloride	Mouse	Lung	Liver	ca. 500
	Mouse	Mammary		
	Rat	Kidney		
	Rat	Liver		

[a] relative tumor rate = $\frac{\text{tumor incidence predicted from most sensitive animal species}}{\text{tumor incidence observed in humans}}$

SOURCE Adapted from National Research Council (262)

from unity by a factor of 3 or less which is assumed for doses scaled on the basis of mg/kg/day (see table 33).

A decision about whether the reasonably good agreement between extrapolated values and observations (72,262) are of significance, and whether or not these findings mean extrapolation is accurate enough for making quantitative decisions about human risks depends on the observer. The NRC Committee **(262)** concluded that:

> Although there are major uncertainties in extrapolating the results of animal tests to man, this is usually the only available method . . . Despite the uncertainties, enough is known to indicate what dependencies on dose and time may operate and to provide rough predictions of induced cancer rates in human populations.

Regulating Pesticides, a report prepared by the National Academy of Science's Committee on Prototype Explicit Analyses for Pesticides (267), **says** that seven previous National Research Council reports have recommended extrapolating from animal data to projected human risk, The Committee (267) took a different position and recommended that only epidemiologic data be used to estimate human risk:

> OPP [Office of Pesticide Programs, EPA] should abandon its attempts to produce numerical estimates of the effects of the use of pesticides on human mortality and morbidity except when reliable human epidemiological data are available. In the usual case, in which major reliance has to be placed on the results of bioassays, those results should be used to construct indicators of the relative pathological activity of the pesticide under review in comparison with other pesticides and compounds.

Documented differences between the metabolism of a chemical in test animals and in humans would be very useful in any attempt at biologic extrapolation. Poor understanding of comparative biochemistry hampers research in the basic biology of cancer. As research con-

tinues, knowledge of metabolic differences between animals and humans may provide clearer direction, but such information will always be limited by strong constraints on studying metabolism of carcinogens in humans.

SUMMARY

Animal tests or epidemiologic studies yield data that relate cancer (or tumor) incidence to exposure levels (dosage) of the substance under study. The accuracy of the relation between exposure and incidence is always limited. Practical restraints on the number of animals that can be tested means that the data are **always subject to significant experimental error; it also means that only relatively high incidence almost always greater than 10 percent, can be measured in the experiments. Epidemiologic studies may be limited by small numbers of people available for study, or by unknown or uncertain exposure levels.** In all cases, deficiencies in experimental design and execution may further limit the accuracy of relating incidence to dose.

Quantitative extrapolation begins with the experimentally determined relationship between incidence and exposure and may use one of several methods to derive an estimate of incidence at exposure levels likely to be encountered in the environment. When animal data are used for extrapolation one of four scaling factors can be used to extrapolate from animal results to expected human response. The scaling factors vary some fortyfold in the risk they project for humans, and agreement has not been reached about which one is most appropriate.

There is also no agreement about which mathematical models best extrapolate from the exposure levels measured in studies to those encountered in the environment. Linear models, which assume that incidence is proportional to exposure at low-exposure levels, are used by Federal agencies. Some other organizations favor nonlinear models in which estimates of incidence decrease faster than dose decreases. A special feature of some models is the incorporation of a threshold, a low, but nonzero exposure level at which the estimated incidence is zero. Nonthreshold models, which are used by Federal agencies, associate some positive estimate of incidence with all doses above zero.

Suggestions are frequently made that careful inspection of available data and testing various extrapolation models against them will allow selection of the best model. Unfortunately data are not sufficient to make such judgments. Another method to decide which model is appropriate is to make projections from animal data and compare those to observed incidence in humans. The cases where human data are available to make comparisons are few, but the conceptually simple, linear, nonthreshold model is reported to estimate human incidence reasonably well.

The increasing importance of short-term tests has led to efforts to extrapolate from them to estimates of carcinogenic risks in humans or animals. Qualitatively short-term tests perform well in predicting whether a substance will be carcinogenic or noncarcinogenic in an animal test. Quantitative agreement between mutagenic potency in short-term tests and carcinogenic potency in animal tests for carcinogenicity is not nearly so good.