Rejoinder

D. A. Freedman and H. Zeisel

INTRODUCTION

The debate is about the scientific foundations of risk assessment and its proper role in policy analysis. Before responding to the discussants in detail, we will try to summarize the points at issue. Consider, then, two evaluations: i) Risk assessment is imperfect, like any other science. Although most of the main steps are fairly well based on objective evidence, in some cases solid proof is not available yet, and then reliance is placed on the consensus of informed opinion. ii) Risk assessment has some points of contact with scientific reality, but the main tenets are based on conjecture.

These two evaluations cover a fairly broad range. We locate risk assessment near one end; its defenders are trying to move it toward the other. The stakes are high, because if the second description is right, and perceived to be right, then risk assessment becomes difficult to use in policy analysis. That may explain some of the tension in the present exchange.

Both the quantitative and the qualitative extrapolation must be considered. On the first, the numerical extrapolation of risk from high doses in lab animals to low doses in humans, there is a fair measure of agreement: its basis is weak. The evidence runs against the current models, and the discussants seem to accept most of the arguments in Sections 3 and 4 of our paper.

The controversy is more about the qualitative extrapolation—the idea that animal carcinogens should be considered as human carcinogens. We remain sympathetic to this idea, but find the evidence for it unsatisfactory. There are three main lines of argument.

(i) *The four-fold table*, which compares animal experiments and epidemiological studies:

		Carcinogenicity in rodents	
		Yes	No
Carcinogenicity in humans	Yes		
	No		

Some of the discussants criticize the way we drew up this table, and we will answer. For their part, the discussants offer no data to show that animal experiments and epidemiology lead to similar classifications of carcinogens; one discussant goes so far as to argue that in principle no such data can exist.

(ii) *All human carcinogens are animal carcinogens.* This is the main argument used to defend the qualitative extrapolation. The discussants, however, admit two of our objections: i) the argument turns the question on its head; ii) with any substance considered carcinogenic in humans, an extraordinary effort goes into finding a sensitive species and test conditions under which carcinogenicity can be demonstrated in the lab, so there is an element of circularity in the reasoning.

(iii) *Biology.* The discussants cite biological findings, but this evidence does not support their positive claims about risk assessment, as we will show. There remains a large gap between risk assessment practice and scientific knowledge. This is so both for the quantitative extrapolation, and the qualitative.

Our own view is that in order to make even moderately reliable risk estimates for human populations, human data are essential. That means epidemiology. Furthermore, in order to make progress in understanding the mechanisms of cancer, for prevention and therapy, laboratory work is needed. The field of molecular biology is advancing toward these goals. Animal experiments are involved here, but they are different in conception from regulatory bioassays. Nor do the models of carcinogenesis now used in risk assessment play a serious role. In the future, of course, models that take into account the molecular biology might prove to be extremely useful.

Even without a better understanding of the mechanisms of carcinogenesis, there might be ways of validating the extrapolation, although these would be slow and expensive: i) doing careful epidemiological studies and bioassays on a test set of substances for which there is already human exposure, and carcinogenicity is suspected; ii) doing animal experiments on a sample of chemicals and species, to measure interspecies differences—for example, can one extrapolate from rodents to primates? There does not seem to be much work along these lines, which amount to systematic versions of the four-fold table mentioned earlier.

Some of the discussants seem unwilling to accept the limitations of risk assessment. Even if the scientific foundations are weak, they seem to be saying to us, risk assessment cannot be worse than the alternatives: what do you propose to do instead? This is a

reasonable question, although not the primary one we are addressing. The answer given in the paper is somewhat woolly; that may be a necessary characteristic of any sensible proposal. But defending risk assessment as better than nothing leads inevitably to holding risk assessment out as more than it is. By pretending to offer knowledge where none exists, risk assessment distorts the policy process and the research agenda.

BRESLOW

After summarizing his objections to our arguments, Breslow goes on to say, "In spite of all this, I share in large measure their skepticism about the scientific value of routine risk assessments that use statistical models fitted to limited animal data obtained at high doses to predict the human response at low ones." So our conclusions are right, but for the wrong reasons. We like having his support where it counts, but our arguments weren't that bad.

The Nature of Risk Assessment

"... risk assessments generally are not viewed by their protagonists as science in the usual sense, but rather involve decision making in the face of uncertainty." This begs the question. Before using a statistical procedure to help make decisions, its operating characteristics have to be assessed. That is part of the job of a statistician. If a statistical procedure turns out to be unreliable, or its reliability is in principle unknowable, it should be given appreciably less weight.

A parable: Some friends are lost in the woods and using a compass to find their way. If the compass is broken, pointing that out is a positive contribution to getting home safely. The reply—"but we're making decisions in the face of uncertainty"—shows some failure to grasp the situation.

The Molecular Biology

We cited some of this evidence, although the field is in rapid progress and more keeps coming in (for updates, see Ames, 1987; Klein, 1987; Varmus, 1987). As we read the literature, it supports the idea that on the way to malignancy, a cell must undergo two or more heritable changes. Armitage and Doll showed good judgment on that score among others. If we did not praise the "biological model" sufficiently, let this make amends.

However, many technical assumptions are needed to move from the qualitative insight to the quantitative dose-response models of risk assessment (Section 3). For example, in the multistage model, the number of target cells is fixed; there is only one pathway to cancer, with a fixed order of stages; and no clonal expansion until the end. We think the molecular biology runs against these assumptions. If the steps to malignancy need not occur in order, and there are multiple pathways, and clonal expansion of intermediate cells matters—the points Breslow makes—then the Armitage-Doll model is wrong. If so, risk assessments based on that model may not be so reliable. The arithmetic cannot provide its own justification.

Perhaps in consequence, when arguing the molecular biology, Breslow attempts to link the "multistage" and "two-stage" models. Even if one has to be abandoned, the other will still be available, and the two sound the same. However, there are substantial differences between them. In the two-stage model, by contrast with the multistage, there is a population of normal target cells. This population is allowed to grow or die. There is a conversion rate for cells from normal to an "intermediate" stage. The population of intermediate cells can also grow or die—"clonal expansion"—and each intermediate cell converts to malignancy at some rate.

Such features allow the two-stage model to fit data for retinoblastoma or Wilms' tumor, the examples given by Breslow in support of the multistage model. As it happens, these cancers are diseases of childhood and cannot show the dramatic increase in age-specific incidence rates demanded by the multistage model. From our perspective, they are clear examples of failure in the latter model, rather than success. In the end, we do not think the multistage model squares with the biology.

While defending statistical models, Breslow seems to compare Mendelian genetics and risk assessment. However, some distinctions are worth preserving. Mendelian genetics is great science because it suggests a variety of experiments and makes sharp predictions about the outcomes, which can be verified, and the mechanisms behind the statistical regularities can be uncovered, leading to a whole series of spectacular advances, both theoretical and practical.

The dose-response models now used in risk assessment cannot be said to have either characteristic. On the other hand, much of the contemporary work in molecular biology, including the work on retinoblastoma and Wilms' tumor, has the mark of excellence. Good models may yet be built on that work.

Abbott's Formula

We are not pushing this formula, but reporting that it is widely used in risk assessment. In standard notation, the conventional multistage dose-response model is

$$P(d) = 1 - \exp\left\{-\sum_{i=0}^{n} a_i d^i\right\}. \tag{1}$$

As in our paper, let

$$(2) \qquad Q(d) = 1 - \exp\left\{-\sum_{i=1}^{n} a_i d^i\right\}.$$

Clearly, $Q(0) = 0$ and

$$(3) \qquad P(0) \doteq 1 - \exp\{-a_0\}.$$

Now

$$(4) \qquad P(d) = P(0) + [1-P(0)]Q(d).$$

That is Abbott's formula, just like we said. There is more to this than algebraic juggling. In many applications, risk is extrapolated from the worst site in the animals to all sites in the humans, which is how the formula comes in and makes a difference (the end of Section 2).

Breslow does not seem to like this extrapolation; we do not like it; maybe they should stop doing it.

Qualitative or Quantitative?

Breslow accuses us of "blatant distortion," on the grounds that our Tables 6 and 7 "seem to imply that a chemical classified as having 'limited' or 'inadequate' evidence of carcinogenicity should be regarded as somehow less carcinogenic than one for which there is 'sufficient' evidence of carcinogenicity."

It seems possible to defend the idea that size of an effect has something to do with strength of evidence, but we do not need to go that far. We always thought grade of evidence meant grade of evidence, not size of carcinogenic effect. The tables, as the captions and column headings make clear, are about such grades or degrees of evidence, and how coherent they are for different species.

What about the substance of the issue? Breslow may believe that if all the evidence were in, with few exceptions the chemicals that are carcinogenic in animals are carcinogenic in humans and vice versa. However, this belief is not grounded in evidence. He even says that explicitly—the data are not there. (That they cannot be developed seems too strong to us; Breslow may be overly pessimistic about epidemiology.)

Our conclusion from the IARC data (end of Section 5): "the research reports of the cancer community (even taken at face value) do not sustain the conventional arguments for the validity of the qualitative extrapolation." Despite all his protestations, Breslow comes rather close to taking the same position.

The DDT Data

In principle, Breslow has a good point: if a chemical is too toxic, it kills the animals before they have time to develop tumors. Our Table 9 does not adjust for life-span and that could be a problem. However, as far as we can tell, life-span does not really affect our argument. To see why, just delete the seven experiments with a significant decrease in life-span ($Z \geq 2$) and recompute. In this smaller group of bioassays, DDT is not life-shortening. It does cause liver tumors and lung tumors in mice, and perhaps thyroid tumors in other rodents (Tables 9 and A). But on balance, it still looks protective at the sites other than the liver (Table B).

The point is about bioassays not DDT. Even after taking life-span into account, many chemicals on test only seem to move tumors from one site to another, rather than increasing total tumor yield. This may be inevitable, given the high background tumor rates in control animals (Table 2). Usual practice in risk assessment is to extrapolate from the site with the largest increase, and ignore all the decreases. That cannot be right.

We make no claim for priority. Salsburg (1983) gets some credit. An even more piquant citation, given the position he took on this issue as a discussant, is Haseman (1983). That paper surveyed 25 NTP bioassays on the Fischer 344 rat with various chemicals on test. The treated animals lived a little longer than the controls, so toxicity was not the problem; but increases in cancer rates at one site were balanced by decreases

TABLE A

A study of studies, recomputed: the impact of DDT and its metabolites on mice, rats and hamsters. Z-tests for dose response in death rates and tumor incidence rates by site. Experiments with significant life-shortening in the dose group are censored.

Z-values	Deaths	Liver	Lungs	Lymphoma/ Leukemia	Osteoma	Kidneys	Testes/ Ovaries	Mammaries	Pituitaries	Adrenals	Thyroid
+2.0 or more	0	14	5	4	0	0	1	0	0	1	0
+0.1 to +1.9	12	6	7	4	0	1	2	2	3	2	5
0.0 exactly	1	6	3	4	1	1	2	2	3	1	2
−0.1 to −1.9	10	0	5	6	5	3	6	5	1	2	0
−2.0 or less	1	0	2	5	0	0	0	1	0	1	0
???	3	1	5	4	21	22	16	17	20	20	20

TABLE B

Z-statistics for dose response in tumor incidence rates by site other than the liver. Mice, rats and hamsters combined; DDT and metabolites. Experiments with significant life-shortening in the dose groups are censored.

+2.0 or more	11
+0.1 to +1.9	26
0.0 exactly	19
−0.1 to −1.9	33
−2.0 or less	9
No data	145
Censored	63

elsewhere. There is something quite odd going on, as Breslow more or less concedes. The discussants should be working on the problem, rather than attacking messengers who repeat bad news.

Other Points

(i) Breslow does not like the review of Tomatis (1979) because "we are not provided with the denominators." Hiding denominators is bad; but the base for our percentages was the set of 26 chemicals under discussion, as the paper makes clear. The idea was to show the incompleteness of the argument that "all human carcinogens are animal carcinogens." Breslow probably wanted us to use other denominators, namely, the subset that were properly tested in animals. That is a reasonable alternative, although the necessary data may not be available.

From our perspective, however, the alternative denominators adjust even less well than ours for a major confounder: the disproportionate experimental effort spent in proving that human carcinogens are also animal carcinogens. We commented on this point in the paper (Section 5.4); Kaldor and Tomatis confirm it. Negative results may not be followed up and published; or eventually some investigator may find an experimental setup in which the substance in question causes cancer in some type of lab animals, and then a long series of published negative reports may be supplanted by a positive finding.

One example: Wilbourn et al. (1986, page 1859) flatly assert that lung cancer can be produced in mice and rats by inhalation experiments, despite numerous failures by many investigators to produce such an effect. For a brief summary, see Doll and Peto (1981, page 1215); for a review of the early literature, see U. S. Public Health Service (1964, page 165 and following) or Wynder and Hoffman (1967, Chapter XII); on the more recent literature, IARC (1986, pages 127 and following).

(ii) Breslow recommends Wilbourn et al. (1986) as an antidote to our Tables 6 and 7. We cited this paper as giving the bioassayist's side of the story, but not claiming the ability to make site-specific predictions from animals to humans. Probably we should have commented on the positive finding and do so now.

The authors say, "for many exposures causally related to human cancer, there is a target organ in common between humans and at least one animal species, despite many inherent physiological differences." Given the variety of experiments and species, this does not seem like a strong claim, especially since Table II in their paper only reports on chemicals carcinogenic to humans.

Now the data. There is some overlap between target sites in humans and in the test animals; but the discrepancies are more prominent. And by our count, 58 of the 94 human target organ responses listed by Wilbourn et al. are starred in the table, meaning that there is a "suspected association" between responses at that site and the exposure. In other words, there is some overlap between *demonstrated* target organs in the animals and *suspected* target organs in humans. Furthermore, there is some overlap between the group that wrote (Wilbourn et al., 1986) and the group that did the suspecting (Merletti et al., 1984):

The writing group: Wilbourn, Haroun, Heseltine, Kaldor, Montesano, Partensky and Vainio.

The suspecting group: Merletti, Heseltine, Saracci, Simonato, Vainio and Wilbourn.

(iii) "Freedman and Zeisel fail to mention one of the more cogent arguments in favor of using linear extrapolation. . . ." We do not see the relevance of low-dose linearity to the extrapolation from data collected at high dose: where is the linearity supposed to start? Figure A represents the issue graphically. Both models are linear at low dose and fit the data equally well but have quite different public health implications.

DuMOUCHEL

A paraphrase of his section headings: He expected not to like the paper; but in his heart, he knows we're right. Although he is on our side, the rules of the game require critical comment. Here goes. DuMouchel writes:

> "DuMouchel and Harris (1983) use a Bayesian model and estimates of prior uncertainty to estimate a human dose-response slope for the risk of lung cancer from exposure to diesel emissions. If the estimation is based on just two animal studies and one human study of a similar chemical, its interval of posterior uncertainty spans six orders of magnitude! Yet, when a set of 37 studies, covering five different biological systems and ten somewhat similar chemicals, is integrated into an analysis using the same basic prior distributions,

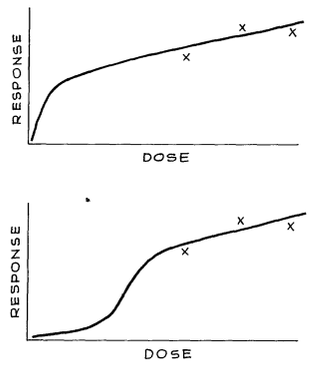

FIG. A. *Low-dose linearity does not help in extrapolation from high dose to low dose. The crosses are "data"; the curves are "models."*

the interval of uncertainty for the same parameter spans just two orders of magnitude, a useful result."

The analysis rides on an assumed linearity of dose response, as the authors note. The six orders of magnitude and the two orders are uncertainties about a parameter in a model. Yet, and DuMouchel knows this better than most of us, the form of the dose-response function is the dominant source of uncertainty when extrapolating risks (especially when most of the data concern not animals but cells on a Petri dish). That uncertainty has not been taken into account. The charm of the quantitative results hid the real problem, for one crucial instant, from the modeler himself. Models sometimes do have a way of sweeping variance under the carpet. DuMouchel not only objects to this point, but illustrates it.

HASEMAN

The Extrapolation from Rodents

Haseman argues that "the primary basis for using laboratory animals such as rats and mice is scientific, not pragmatic." For evidence, he refers to Freireich et al. (1966), Rall (1979), Huff and Moore (1984) and some official literature. Arguments from authority no longer impress as much as they used to, so let us consider the evidence in the references.

Freireich et al. (1966) are writing about toxicity not carcinogenicity. That seems irrelevant here.

Rall (1979) begins with the ex cathedra pronouncement quoted by Haseman, but goes on to make an argument for the validity of risk assessment. The main evidence consists of an unpublished manuscript whose data base is identified as the IARC Monographs 1 to 13, a subset of the data we have been using. There are two analyses. The first is rather like our Table 7, but reaching the opposite conclusion; every single one of the 81 animal carcinogens turns out to be carcinogenic in humans, although as late as 1982, the IARC could only positively identify 30 such compounds. The second point is not unfamiliar: all human carcinogens are animal carcinogens. Rall goes on to cite some authorities for more support. One is the IARC. Another is the National Academy of Sciences (1975) report customarily mentioned–along with Rall (1977)–by the IARC when it wants to justify risk assessment.

Huff and Moore (1984) begin by reciting the argument that all human carcinogens are animal carcinogens. They then appeal to authority, including Freireich, Huff and Moore, the IARC, the National Academy of Sciences, the Office of Technology Assessment, and Rall. That seems to be all the evidence behind the passage reproduced by Haseman, and the names are beginning to sound familiar.

The emotional tone in Huff and Moore is also worth noting:

> "History remains clear: all known human carcinogens cause cancer in animals. Should society not demand therefore that until absolutely and unequivocally proven otherwise unnecessary exposure to all convincing animal carcinogens must be eliminated or substantially reduced? Dreadfully the 1000 cancer deaths per day in the United States alone testify that better and more complete preventive measures are justified. Even if one adopts the conservative and lower estimate of cancers definitely caused by chemical exposure this represents 20,000 human deaths minimum per year attributed directly to chemicals per se. Further, other chemically-induced subtle or clinically manifest toxicities must be viewed with equal concern."
>
> "Given the 6,000,000 unique chemicals cataloged in the Chemical Abstracts system, given the 70,000 chemicals in regular use, given the 1,000 to 3,000 chemicals introduced annually into the market place signal a formidable toxicological task ahead."

What we see is a group of authors who firmly believe in the need for bioassays and the validity of risk assessment. They cite each other for support. We respect the beliefs, but not the evidence. Indeed, there are clear examples of strain or species differences in response to carcinogenic insult, and sometimes the biochemistry is understood. Two elegant examples: Boutwell (1964) and Rossi et al. (1983). The list could be extended (Cairns, 1978, pages 97–104, 129–32; Wald and Doll, 1985). Species differences do exist. Before relying too much on bioassays to screen for carcinogenicity, these differences need to be quantified.

To conclude the discussion on the scientific basis for using rodents in bioassays, we want to quote an

authority of our own, Tomatis (IARC, 1980, page 30):

> "The species more frequently used in carcinogenicity studies are rats, mice and Syrian hamsters. The rat and mouse have been used more extensively than the hamster, but the latter is particularly suitable for studies on respiratory-tract carcinogenesis [references omitted]. The wide use of these three rodent species, which are still the species of choice for carcinogenicity testing, is not based on biochemical, physiological or anatomical similarities between these animals and man, but rather on practical considerations such as relatively short life-span, small size and availability."

The Correlation between Rats and Mice

Haseman chides us for "selectively ignoring other more comprehensive studies. One of the studies [Freedman and Zeisel] chose not to discuss is Purchase (1980).... Rather than citing this more definitive study, Freedman and Zeisel instead attempt to carry out their own evaluation of two subsets of the carcinogenicity studies carried out by the National Cancer Institute (NCI) and the National Toxicology Program (NTP)."

Now we do cite Purchase (1980), right there in the bibliography; and even discuss the paper (Section 5.2), saying that it comes out on the other side of the argument. However, we do not give it much weight, and here is why. Purchase draws his data from three different sources: The NCI/NTP experiments, the IARC monographs and "the literature," that is, all others. But the quality of "the literature" is quite variable. And individual studies aside, there is a real publication bias—negative findings tend not to get published. If a reference is needed, see IARC (1982, page 12) quoted below; they are to be complimented on their candor.

The IARC assessments used by Purchase come down to a set of committee reports on whether certain chemicals should or should not be classified as carcinogens for various species. Decisions are based on literature reviews, and are reached by consensus, sometimes by voting. So Purchase only demonstrates the obvious: IARC working groups tend to think on the whole that substances carcinogenic in the mouse are carcinogenic in the rat (and the human too, if the truth were known).

The NCI/NTP really is the best data set for present purposes. There is a fairly uniform protocol for the studies, and it is quite good, thanks in large measure to the work of Breslow and Haseman. Furthermore, the experiments almost always involve both rats and mice, so there is a series of paired comparisons. Finally, publication bias is minimal.

Why does Haseman object so strongly? He did the same sort of analysis himself (and very nicely too) in Haseman et al. (1984). Aren't lawyers and statisticians allowed to count?

Another charge is that we only used "subsets" of the data, rather than Haseman and Huff (1987). In fact, we used all the data reported by our sources. And none of this can really matter, because our two "subsets" look just like the most recent data; the breakdowns of the NCI/NTP data are quite stable over time. For example, consider the percentage of compounds positive in both rats and mice, among those positive in at least one species. The three studies compare as follows:

Griesemer and Cueto (1980)	45%
Haseman et al. (1984)	40%
Haseman and Huff (1987)	50%

So what is the real issue? In his 2×2 table, Haseman is impressed by the numbers on the diagonal. We are struck by the numbers off the diagonal. That's all there is to it. And our original summary still seems fair (Section 5.2):

> "There is no clear bottom line to report. Taking all the experiments at face value, there is some measure of agreement between the results for rats and mice, and some measure of disagreement. Now rats and mice are much more similar to each other than either is to humans. The validity of the mouse-to-man extrapolation seems hard to argue on the basis of these data."

The Correlation with Human Data

Our critics all attack Table 7, but cannot agree on just why it is so bad. Breslow seemed to think that Table 7 was about carcinogenic potency; Haseman knows it is about the qualitative correlation, but rejects the row and column definitions, because "it is INAPPROPRIATE to combine 'limited,' 'inadequate' and 'no evidence' categories for purposes of statistical analysis."

To begin the APPROPRIATE response, Haseman may not have the definitions quite right. Now it is our turn to quote from the IARC (1982). The categories for human evidence are on page 11:

> "The degree of evidence for carcinogenicity from studies in humans were categorized as:
>
> i. *Sufficient evidence* of carcinogenicity, which indicates that there is a causal relationship between the agent and human cancer.
> ii. *Limited evidence* of carcinogenicity, which indicates that a causal interpretation is credible,

but that alternative explanations, such as chance, bias or confounding, could not adequately be excluded.
iii. *Inadequate evidence*, which indicates that one of three conditions prevailed: (a) there were few pertinent data; (b) the available studies, while showing evidence of association, did not exclude chance, bias, or confounding; (c) studies were available which do not show evidence of carcinogenicity."

The prose is a little dense, but the idea is clear: There is no category iv, and negative studies are classified as *inadequate*. In particular, bioassayists like Haseman who want to eliminate category iii from their tabulations also want to eliminate all the evidence against their position, because category iii contains all the negative studies on carcinogenicity in humans.

The parallel definition for the animal evidence should also be available in full (IARC, 1982, page 12):

"i. *Sufficient evidence* of carcinogenicity, which indicates that there is an increased incidence of malignant tumors: (a) in multiple species or strains; or (b) in multiple experiments (preferably with different routes of administration or using different dose levels); or (c) to an unusual degree with regard to incidence, site or type of tumour, or age at onset. Additional evidence may be provided by data on dose-response effects, as well as information from short-term tests or on chemical structure.
ii. *Limited evidence* of carcinogenicity, which means that the data suggest a carcinogenic effect but are limited because: (a) the studies involve a single species, strain, or experiment; or (b) the experiments are restricted by inadequate dosage levels, inadequate duration of exposure to the agent, inadequate period of follow-up, poor survival, too few animals, or inadequate reporting; or (c) the neoplasms produced often occur spontaneously and, in the past, have been difficult to classify as malignant by histological criteria alone (e.g., lung and liver tumours in mice).
iii. *Inadequate evidence*, which indicates that because of major qualitative or quantitative limitations, the studies cannot be interpreted as showing either the presence or absence of a carcinogenic effect; or that within the limits of the tests used, the chemical is not carcinogenic. The number of negative studies is small, since, in general, studies that show no effect are less likely to be published than those suggesting carcinogenicity.
iv. *No data* indicates that data were not available to the Working Group.

Again,the point is clear: animal studies that show a compound to be noncarcinogenic are classified as *inadequate*.

Back to Table 7. As we understand the source for the data (IARC, 1982, page 8), the starting point was a set of 585 chemicals, groups of chemicals and processes. The objective was to study the ones with human data, and that lead to the 155 listed in Table 1 of the source document. (A spot check does indicate that in roughly 20% of the cases, the human data are sketchy or nonexistent; that may be the point discussants are trying to make.)

We eliminated 19 cases where the IARC said that animal data was irrelevant, and the 3 with no animal data, leaving the test set of 133 in Table 7. From our perspective, these are the chemicals most likely to prove carcinogenic in humans; that is why they were studied so carefully. Still from our perspective, *limited* or *inadequate evidence* ranges from almost nothing to strongly negative, both for the animals and for the humans. That is why we combined the categories as we did. If our perspective seems unreasonable, please reread the definitions.

Obviously, Haseman is free to tabulate the data any way he wants. But our tabulation is just as APPROPRIATE as his, if not more so. Interestingly, what he does with his tabulation is to rush off and compute the fraction of human carcinogens that are animal carcinogens; he even does this a second time in his closing section; Kaldor and Tomatis insist on doing it too. Evidently, this is a calculation with a lot of emotional horsepower. But it addresses the wrong question; and maybe they even get the wrong answer, because they ignore the data that doesn't fit (see Table 6 and discussion).

Haseman's punchline. "Clearly, Freedman and Zeisel's conclusion that the IARC data summarized above provides 'decisive evidence' that 'carcinogenicity in lab animals is poor evidence for an effect in humans' is a misrepresentation of these data." The misrepresentation is of our position. We know there are plenty of false negatives (and false positives) in the data; that chemicals get selected for testing in quite odd ways; and exposure is an important issue. So here is what we really said (end of Section 5, emphasis supplied):

"*In principle*, the evidence in Table 7 is decisive. Carcinogenicity in laboratory animals is poor evidence for an effect in humans. *Questions about the representativeness of the test set and doubts about the quality of the underlying studies (both positive and negative) weaken this conclusion appreciably.*"

And on the real issue, our (weak) conclusion still seems right:

> "If a substance is carcinogenic in a bioassay, we think that is some evidence for carcinogenic potential in humans. If the bioassay was well run, the evidence is stronger. Replicability across experiments and across species makes the case even stronger. Conversely, flaws in the experiment or failure to replicate weaken the argument."

Recommendations

Not surprisingly, Haseman rejects most of our recommendations for improving bioassays.

a. *Randomization.* Haseman's counteroffer (the NTP 1984 protocol) is fine with us. We wish more people would follow his advice.

b. *Pooling.* Haseman says "Pooling a hodgepodge of biologically unrelated tumors does not result in a meaningful variable for biological or statistical analysis." Different investigators recognize different categories of tumors, and even the NCI/NTP (which has the clearest and most consistent definitions) changes its mind from time to time. In other words, the categories Haseman is defending result from committee meetings rather than the laws of biology. He should not resist alternative proposals quite so vigorously.

If we combine the bioassayists' arguments on the qualitative extrapolation and on pooling, the position really comes down to this: all species of animals are quite similar, but currently recognized rodent tumors are all quite different. For the quantitative extrapolation, they would have to go further. It is a standard practice to select the most sensitive species, sex and site in the lab animal, and extrapolate the increased rate of tumors at that one site to all cancers in humans. They pool the whole "hodgepodge of biologically unrelated" human cancers, but refuse to pool in the rodent.

c. *Defining end points.* "It makes no sense to discount an obvious site-specific carcinogenic effect merely because the investigator was unable to specify the target organ and tumor type in advance." People who do experiments should keep their eyes open, true enough. But if they change their minds about the rules of evidence as they go along, the P-values do not mean much. Replication would help more than simulation studies on multiple testing, because the ingenuity of the bioassayist is hard to capture in the computer code.

The comment that "it is simply not possible in most cases to predict precisely what carcinogenic effects will be observed in laboratory animal studies" seems right on the money—even after several experiments on the chemical have been completed. Nor is the qualifier "precisely" really needed. So why does anyone think such lab data can be used to predict results for humans?

d. *Blinding.* The advantages still seem to outweigh the drawbacks.

Other Issues

Lawyers. Haseman fears harm to public safety when lawyers get near the regulatory process. If dispassionate scientists are making public policy on the basis of objective fact, who needs lawyers? However, the present exchange shows the hypothesis to be an attractive fantasy. The real game looks more like advocacy disguised as science.

Do they or don't they? What we said: "Because there are a variety of standard randomization schemes, we lean to the view that the other authors [whose papers don't discuss randomization] did not, in fact, randomize the animals to the various dose groups." What Haseman says: "Throughout the paper Freedman and Zeisel display an arrogant attitude toward nonstatisticians, (e.g., assuming that investigators do not randomize properly unless the randomization scheme is stated explicitly . . .)."

At first reading, he seems to be saying i) we have a bad attitude, and ii) they really do randomize. Read it again. He is only claiming i. The reason he does not claim ii is obvious.

Misleading interpretations. Haseman accuses us of a "misleading interpretation" of the evidence on the lung in Haseman and Hoel (1979) because our summary, although "technically correct," implies greater variance in the risk estimates than Haseman and Hoel were describing.

Now we cited four of his papers, and only got one of them half-wrong. He is a tough grader, so 3½ might be acceptable, but we want 4/4. The way to get it is just to quote the disputed passage from Haseman and Hoel (1979), which is worth reading anyway.

> "As a result of these differences, human risk estimates varied widely (Table 7), ranging from no discernible risk to 5.3×10^{-5} Q. [Q is the dose rate, in $\mu g/d$.] *Even for studies in which some increased risk was indicated, there was as much as a 1000-fold difference in the actual risk estimates for lung tumors. For lymphomas and liver tumors, the agreement was better, particularly for the males.*"

> "The DDT data showed evidence of study-to-study as well as strain-to-strain variability. The 9-fold increase in lymphosarcomas produced by DDT found by Tarjan and Kemeny was not observed by Terracini et al. (1973), who used the same inbred strain and doses ranging from essentially that used by Tarjan and Kemeny to nearly

100 times as much. The 14-fold increase in pulmonary carcinomas observed by Tarjan and Kemeny was also not seen by Terracini et al. The control rates of pulmonary adenomas were also markedly different in the two studies (5% compared with 28–42%)."

"Other inconsistencies could be cited: the marked differences between sexes in the incidence of liver tumors observed by Turusov et al. (1973) and Tomatis et al. (1972) was not observed by Thorpe and Walker (1973) and Walker et al. (1972), who used the same random-bred strain. For other minor discrepancies, see Tables 3–5. *Although the variability among strains seemed real, the magnitudes of certain differences among studies involving the same strain were somewhat disconcerting. Use of different sublines could partially explain the discrepancies for the inbred strains, and different suppliers could be a factor for the random-bred animals.*" [emphasis supplied]

Sometimes, Haseman takes our side of the argument.

KALDOR AND TOMATIS

We gratefully acknowledge the major points of agreement, for example, "that, in particular, the quantitative assessment of cancer risk entails a number of biological assumptions which have not been verified empirically." The main issues have already been canvassed, so we respond only to a few interesting points of detail.

Quality of Bioassays

Kaldor and Tomatis candidly admit that "Overall, the quality of data from animal cancer tests is very uneven." But, they continue, "Freedman and Zeisel chose to ignore these differences in quality in their presentation [of the DDT data]." As we see it, these quality differences are not a flaw in the analysis, but a primary conclusion. (If Kaldor and Tomatis are hinting that the positive studies were the good ones, they should look at Tables 8 and 9 again.)

Whether these quality differences are properly factored into the IARC evaluations is a harder question. Despite the assurances of Kaldor and Tomatis, the DDT example is troubling. In 1982, the human evidence was judged *inadequate*; in our opinion, it was quite negative. The animal evidence was judged *sufficient*, although the bioassays gave mixed results and some "were certainly deficient in various respects," in the careful words of the discussants. In the end, DDT was rated as "probably carcinogenic to humans," despite the negative epidemiology and the conflicts in the bioassay evidence.

Biology

Like Breslow, Kaldor and Tomatis appeal to recent work in molecular biology. For example, they stress measurement of the "effective dose" delivered to the target tissue, rather than the "applied dose" in food, water, etc. That is fine as far as it goes, but the literature is in some disarray on dose-response relationships even for applied dose (see Brambilla et al., 1987 for one view and Swenberg et al., 1987, for another).

We do agree that if more information becomes available, the extrapolation may get to be on more solid ground. However, Kaldor and Tomatis claim that "Recent observations on the mechanisms underlying liver tumor induction in mice seem to strengthen their significance as predictors of human risk (Reynolds et al., 1987)." This guarded sentence is still an overstatement. Reynolds et al. make a significant contribution to distinguishing spontaneous from chemically induced tumors, by measuring the activity of several oncogenes. But here are the two key passages about the mouse liver and risk assessment:

"The validity of mouse liver tumor end points in assessing the potential hazards of chemical exposure to humans is a controversial but important issue, since liver neoplasia in mice is the most frequent tumor target tissue end point in 2-year carcinogenicity studies."

"Information at a molecular level will make the results from rodent carcinogenesis studies more relevant to the assessment of human risk."

The first sentence is in the present; the second, in the future.

Qualitative Correlation of Carcinogenicity

"Thus, discarding chemicals for which there is insufficient data, the sensitivity of animal experiments in detecting human carcinogens approaches 100%. This figure contrasts sharply with the 59% arrived at by Freedman and Zeisel, using rather selective criteria (Table 6 of their paper)." Wait a minute; this is the old argument that all human carcinogens are animal carcinogens. Our Table 6 lays out all the data; it's Kaldor and Tomatis who get their effect by being "selective."

In connection with our Table 7, they write "It is completely inappropriate to equate an absence of data with a clear negative finding. This error appears in other places in the paper, including the negative classification of experiments quoted in Table 5." We already defended Table 7—in depth—against a similar attack by Haseman.

Table 5 uses data from Tomatis, Partensky and Montesano (1973, Table I). Those authors are trying

to show that results in the mouse predict results in the rat and the hamster, so the extrapolation to humans may be feasible too. In the aggregate, their numbers look quite good. Then we took the class of chlorinated hydrocarbons. For those chemicals, our table shows that on the rat, the data go the other way; on the hamster, there are no data except for one lonely counter-example. Absence of data may make the heart grow fonder, but cannot prove concordance.

Furthermore, our table reports chemicals as "positive, negative, not tested," following the usage in Tomatis, Partensky and Montesano (1973, Table I). So Kaldor and Tomatis are just wrong: we did not "equate an absence of data with a clear negative finding."

As it turns out, however, the bioassayists commit this statistical impropriety on a routine basis. Indeed, the IARC changes its definitions from time to time; by 1986, there is a fourth category for the epidemiological studies on carcinogenicity: *no evidence*. (See Wilbourn et al., 1986, Table I.) Hang on to your seats. That doesn't mean missing data. It means "several adequate studies are available which do not show evidence of carcinogenicity."

One more time: *no evidence* means good negative studies, whereas *limited evidence* means shaky positive data.

Bernstein et al. (1985)

Kaldor and Tomatis are confused, like most of us, by missing data. They are also puzzled by our discussion of Bernstein et al. On that, we can help; and will, because the story is an interesting one. It begins with a paper by Crouch and Wilson (1979) showing a strong correlation between carcinogenic potency in mice and rats. However, Bernstein et al. indicate that Crouch and Wilson were misled by a statistical artifact of bioassay design. The essential point is that the test set of Crouch and Wilson consisted of chemicals with a statistically significant potency estimate. (The rejected null hypothesis is that the potency k in the one-hit model equals 0; presumably, Crouch and Wilson wanted to avoid talking about the "potency" of non-carcinogens.)

To explain the artifact, Bernstein et al. introduce a somewhat simplified but realistic bioassay with a control group and a dose group of 50 animals each, who are given the MTD (maximally tolerated dose); not all the animals get cancer. In that setup, if the estimate of potency is statistically significant, it must be on the order of 1/MTD. The MTDs are strongly correlated for rats and mice, and range over many orders of magnitude. The potency correlation follows.

One of us is working with some of Bernstein et al. to see if there is any correlation beyond the artifactual one. There may be, a little; and the absence of 100% response rates may be another artifact or a clue to something interesting.

Quantitative Extrapolation

In effect, Kaldor and Tomatis say they knew the whole story already, although we told it so badly; and the remedy is to continue business as usual. The response could be equally brief: If ignorance is no excuse, neither is knowledge.

One passage, however, seems to call for more extended comment:

> "Scientists are all too aware of the complexities involved to take any model of carcinogenesis literally, but generally accept the need for some sort of standardized quantification of the results from animal experiments" (Peto et al., 1984).

This could be misleading, because Peto et al. (1984) do not take a position one way or another on risk assessment, but introduce the Gold et al. (1984) data base where a series of bioassays are summarized using a relatively new statistic. The idea is only to explain the procedure and its properties.

Furthermore, Peto and coauthors are clearly on our side of the argument, as noted in our paper: Ames (1983); Ames, Magaw and Gold (1987); Doll and Peto (1981, pages 1215–1216). Peto (1985) will be discussed below.

MOOLGAVKAR AND DEWANJI

We have learned a lot from Moolgavkar and his coauthors over the years, in print and in person; so the convergence of views is not completely unexpected. Still, it is a pleasure to acknowledge the debt. These investigators are largely responsible for developing the "two-stage" model of carcinogenesis, which may prove to be quite useful in risk assessment, at least for certain cancers. The assumed dynamics would need to be validated, and the extrapolation of rates from lab animals to humans would also have to be considered.

In any case, the two-stage model is more attractive than its predecessors because it is better connected to the molecular biology. The investigators are also paying attention to the rates at which carcinogens are absorbed and metabolized. That all seems to go in the right direction, as do the comments by Moolgavkar and Dewanji on the multistage model.

Protagonists Off Stage

The *Environmental Protection Agency* uses risk assessment to make policy. As this went to press, the agency finished its own reappraisal (*The New York Times*, January 4, 1988, pages 1 and 11). Many of the numbers changed, some up, most down: the estimated

risk for dioxin was cut by a factor of 16; for arsenic, by a factor of 100. The rationale involved a new generation of models:

> "Dr. Moore, the EPA official in charge of toxic substances, said that while regulatory agencies should 'err on the side of being conservative' when it comes to evaluating the risks of chemicals, particularly those that cause cancer, 'this agency has been identified by many as possibly adopting an ultraconservative approach.' Many of the previous assessments, Dr. Moore said, were based on 'simplistic' assumptions and 'now strain credulity.'"

The new models are described as more "sophisticated." Whether they are more reliable remains to be seen.

Armitage and Doll were not discussants, but their model has been mentioned, so a word about them is in order. We esteem them highly, and doubt their views of the model or risk assessment are much different from ours. Doll has already been quoted several times. Here is Armitage (1985):

> "Until and unless we obtain direct evidence about the presence and nature of intermediate stages, any statistical theory is likely to remain largely unfalsifiable, particularly if it is allowed to be modified with the flexibility to which we have become accustomed."

Richard Peto, like Richard Doll, is a great epidemiologist, and a leading expert on bioassays and mathematical models for carcinogenesis. Peto was cited by discussants as a supporter of risk assessment based on animal data, but in fact he is strongly opposed. Peto (1985) introduces a fictitious character called a "hygienist," who in the absence of reliable human data uses the extrapolation from lab animals. The "hygienist" [for which read *risk assessor*, and we apologize to the real hygienists] finds the epidemiologist a nuisance.

> "Conversely, from the epidemiologist's viewpoint, the hygienist, with a long list of worrisome chemicals, may seem at best an irrelevancy and, at worst, a thoroughly diversionary influence. On our present knowledge, the chief priority for cancer prevention obviously is control of the effects of smoking [citations omitted], yet incessant coverage in the mass media of one real or suspect carcinogen after another makes it unnecessarily difficult to get the public to appreciate the unique importance of smoking. (Indeed, for many years the health warning on saccharin in the United States was at least as strong as that on cigarettes!) Moreover, in the search for knowledge about other major causes of human cancer in the developed countries, the hypotheses that appear most promising to many epidemiologists are perhaps those involving infective, nutritional or hormonal factors, rather than those involving exposure to traces of man-made chemicals. So, even within the world of research, the industrial hygienists' interests may be irrelevant or diversionary to the epidemiologist."
>
> "The fundamental source of divergence may be that the real purposes of many epidemiologists are not the same as those of many hygienists. The essential purpose of epidemiology is to understand (and, if possible, control) some major avoidable cause(s) of human disease. The fundamental purposes of many hygienists are rather different. To be sure, many are indeed chiefly concerned with the alleviation of human disease (although, if so, why don't they put at least as much political effort into the discouragement of smoking or of saturated fat intake as into industrial hygiene?) but others are not: whatever their original motives may have been, some appear to be chiefly concerned either with defending or with disliking modern industrial society. Out of their repeated confrontations is emerging the current approach to risk assessment. Because of their roots in adversarial politics, risk assessment procedures sometimes appear to be designed chiefly to be legalistically defensible rather than epidemiologically sensible . . . ."

CONCLUSIONS

Many workers in the field of risk assessment may feel obliged to use techniques that are not scientifically justified, because they see a clear and present danger—an epidemic of cancer caused by synthetic chemicals, which must be brought under control. Their motive is to do good; but, along with most epidemiologists, we do not accept the major premise. There is no epidemic apart from lung cancer, where the primary cause is smoking. Our critics yield the point, if only by silence.

Peto's views were quoted above. Another respected commentator put it this way (Abelson, 1987, in a *Science* editorial on Cancer phobia):

> "For more than 10 years, the public has been subjected to a media barrage leading to widespread, misinformed fear of chemicals. Through the use of questionable evidence, many major substances have been labeled carcinogens. If data are adjusted to eliminate effects of cigarette smoking, there has been no overall increase in cancer due to other factors. The highly publicized

cancer epidemic that was predicted earlier has not materialized."

Abelson may be a bit rough, but he's right. The scientific basis for risk assessment is quite weak; so is the implicit pragmatic justification. Quantitative risk assessment has been oversold, and it is time to reconsider.

On the other hand, we agree that bioassays can be useful in identifying potential hazards. The real question is about the weight to be put on such results, especially when other kinds of evidence are available. Furthermore, much could be done to improve the quality of bioassays, and to validate the extrapolations; although it may be preferable to put incremental resources into epidemiological studies or basic research in molecular biology.

ACKNOWLEDGMENTS

For comments on various drafts of this rejoinder, we must thank A. Adhikari, L. Bazel, R. Daggett, P. Diaconis, L. Gordon, J. Horowitz, J. Pitman, T. Speed and A. Tversky.

ADDITIONAL REFERENCES

ABELSON, P. H. (1987). Cancer phobia. *Science* **238** 473.

AMES, B. N. (1987). Measuring oxidative damage in humans: Relation to cancer and aging. In *Proceedings of the IARC Meeting on Detection Methods for DNA-damaging Agents in Man.* To appear.

ARMITAGE, P. (1985). Multistage models of carcinogenesis. *Environ. Health Perspect.* **63** 195–201.

BRAMBILLA, G., ET AL. (1987). Dose response curves for liver DNA fragmentation induced in rats by sixteen *N*-nitroso compounds as measured by viscometric and alkaline elution analyses. *Cancer Res.* **47** 3485–3491.

FREIREICH, E. J., ET AL. (1966). Quantitative comparison of toxicity of anticancer agents in mouse, rat, hamster, dog, monkey, and man. *Cancer Chemother. Reports* **50** 219–244.

HASEMAN, J. K. and HUFF, J. E. (1987). Species correlation in long-term carcinogenicity studies. *Cancer Lett.* **37** 125–132.

HUFF, J. E. and MOORE, J. A. (1984). Carcinogenesis studies design and experimental data intepretation/evaluation at the National Toxicology Program. In *Industrial Hazards of Plastics and Synthetic Elastomers* 43–64. Liss, New York.

IARC (1986). *Tobacco Smoking. Monograph* **38**. Lyon.

KLEIN, G. (1987). The approaching era of the tumor suppressor genes. *Science* **238** 1539–1545.

MERLETTI, F., ET AL. (1984). Target organs for carcinogenicity of chemicals and industrial exposures in humans: A review of the results in the IARC monographs on the evaluation of the carcinogenic risk of chemicals to humans. *Cancer Res.* **44** 2244–2250.

PETO, R. (1985). Epidemiological reservations about risk assessment. In *Assessment of Risk from Low-level Exposure to Radiation and Chemicals* (A. D. Woodward et al., eds.). Plenum, New York.

PETO, R., ET AL. (1984). 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.

RALL, D. P. (1979). The role of laboratory animal studies in estimating carcinogenic risks for man. In *Carcinogenic Risks—Strategies for Intervention* (W. Davis and C. Rosenfeld, eds.). *IARC Sci. Publ.* **25**. IARC, Lyon.

REYNOLDS S. H., ET AL. (1987). Activated oncogenes in B6C3F1 mouse liver tumors: Implications for risk assessment. *Science* **237** 1309–1316.

SWENBERG, J. A., ET AL. (1987). High to low dose extrapolation: Critical determinants involved in the dose response of carcinogenic substances. *Environ Health Perspect.* To appear.

UNITED STATES PUBLIC HEALTH SERVICE (1964). Smoking and health. Report of the Advisory Committee to the Surgeon General of the Public Health Service. Dept. Health, Education and Welfare, Washington.

VARMUS, H. E. (1987). Oncogenes and transcriptional control. *Science* **238** 1337–1339.

WYNDER, E. L. and HOFFMAN, D. (1967). *Tobacco and Tobacco Smoke.* Academic, New York.