D R A F T

CARCINOGENIC SAFETY AND THE THRESHOLD CONTROVERSY

(by Paul F. Deisler, Jr.)

Description of the Controversy

Dose-response, or the variation in intensity of effect with level of dosage or exposure, is a well known characteristic of living creatures. It is commonly thought, however, that below some threshold level of dosage, particular toxic agents do "no harm"; certainly, for many simple, acute effects, this appears to be the case. Some illustrative examples of possible dose-response curves are shown in the attached Figure.

The error inherent in test data is usually large, even in well conducted experiments. The response, if calculated as the fraction of the test animals showing at least one adverse effect, is usually taken as an estimate of the average binomial probability, p, of a test animal exhibiting the effect and the data are so treated. Even in a simple, perfectly conducted experiment involving n animals where each has a measured probability, p, of exhibiting an adverse effect at a given dose, there is a variance which is np(1-p); in less-than-perfect experiments, other experimental errors engendered by the practical impossibility of

exactly reproducing all experimental conditions or of excluding bias increase this basic variance. Because of this inherently variable character of test data, even very well done experiments with very large numbers of test animals cannot "prove", precisely and rigorously, that no response will be observed below some greater-than-zero dose level for a given agent. Only probabilistic statements may be made, at best, involving specified confidence levels.

For carcinogens which act directly on DNA, theoretical mechanisms of carcinogenesis can lead to the conclusion that no non-zero dose, however small, is without effect; moreover, reasonable sounding mechanisms such as the so-called "one-hit" mechanism (Curve A) or the multi-stage mechanism (Curves B or C) can approach a linear-through-zero dose-response function as dose approaches zero. Consideration of these possible mechanisms together with consideration of the uncertainties discussed above and of the fact that there are as yet no general pharmacokinetic theories of carcinogenesis which permit the "right" mechanism to be selected for extrapolation from experimental dose levels to the usual low levels of exposure humans might experience, when combined with a strong desire to be conservative, has all led some of the regulatory agencies to conclude that their policy should be to state that any exposure to a carcinogen, however small, must be assumed to increase the risk of cancer and that they must regulate accordingly.

Moreover, some also conclude that some form of linear-through-zero, upper-confidence limit extrapolation should be used in estimating risk, thus running the risk of over-estimating the risk by factors from about two up to factors of several orders of magnitude. In effect, the possibility of a threshold existing is ignored, as a matter of policy, unless perhaps some method is found, in a given case, to clearly demonstrate that one exists. At this point in time it is not known what the attitude of any such regulatory agency would be if confronted with such a demonstration.

Epigenetic carcinogenic mechanisms can be postulated which lead to response-functions which have thresholds (T, Curve D). Also, functions having thresholds or "practical" thresholds can be derived for reasonable pharmacokinetic mechanisms involving various combinations of different types of phenomena: DNA repair, immune reactions, non-direct carcinogenic action via metabolites, reversible reactions, deactivation reactions, competing reactions, phase changes, and so forth. Such a special, theoretical mechanism has been studied by Cornfield.(a) Many shapes of curves can result from such mechanisms. Since latency period is usually also a function of dose, the argument has been

(a) Cornfield, J., "Carcinogenic Risk Assessment" Science, 198, 693 (1977).

made that, at least for some individuals, dose levels exist for which the latency periods exceed the individuals' lifetimes and, consequently, such dose levels are thresholds for those individuals. Moreover, the mere existence of dose-response has been thought to suggest the existence of individual thresholds since, it can be argued, the individuals in an experimental group subjected to a given dose that show responses may be considered as showing that that dose is at or above their thresholds while those showing no response may be considered as showing that that dose is below their thresholds, whatever those thresholds may be individually. This latter reasoning is circular, but, nonetheless, for a combination of the above reasons, some are convinced that thresholds exist, at least in some cases, and that the effort to define them should form an important part of risk assessment. Some definitions of thresholds and further examination of the various arguments are pursued in the next section.

Different Types of Thresholds and Further Arguments

The above represent the poles in a highly polarized situation. One problem exists which helps prevent the participants in the debate from understanding in what way they disagree with each other is the definition of "threshold". The definitions of three types of threshold are here suggested: the individual threshold (I.T.), the absolute population threshold (A.P.T.), and the practical population threshold (P.P.T.).

The individual threshold (I.T.) is the threshold an individual may have for a given agent; such thresholds may differ from individual to individual. Many of the arguments given above for the existence of "thresholds" in fact deal with the possible existence, specifically, of the I.T., though some of the epigenetic mechanisms offer fairly credible support to the idea of the existence of the A.P.T. (as defined further on).

Discussions with knowledgeable people have shown me that, having distinguished between the I.T. and the A.P.T., there is still a diversity of views as to what an I.T. really is. Briefly, two distinct views have emerged, as follows:

1. Each individual exhibits an individual dose-response function such that at a dose, D, there is a probability, p(D), of an adverse effect occurring, and if the dose-response is such that p(D)=0 for some $D=D_T>0$, then that dose, D_T, is the threshold dose for that individual; thus, for doses below D_T no adverse effect will occur to the individual, whereas for doses above D_T an adverse effect may occur and the probability of its occurrence is greater than $p(D_T)$. For this model, the curves in the attached figure may be taken to illustrate four different individual dose-response functions of which only one has a threshold. Dose-response curves derived from samples of dosed subjects

would just be the experimentally measured estimates of population dose-responses which, here, are the weighted sums of the individual dose-response functions. For different populations such estimates of population dose-response curves could exhibit the same diversity of forms illustrated in the figure. For this model individual dose-response is itself, at any dose, D, a distributed function for a population.

2. Each individual exhibits an individual and specific response to dose such that for a specific dose a specific individual either will or will not respond for this model; the dose at which an individual responds is the threshold dose, D_T, for that individual. In this case the probability that an individual will respond at $D>D_T$ is 100%, and a measured dose-response curve is just a measure of the cumulative distribution of the individual D_T-values for a population.

In the further discussions below the first variety of I.T. will be referred to as an I.T. of the first kind and the second, as an I.T. of the second kind.

I.T.'s of the first and second kinds cannot be distinguished from each other by means of the usual, statistical chronic tests, alone. If, for example, in an experiment with n animals the measured value of p is taken as the individual,

binomial probability of the occurrence of an adverse effect, as in the case of an I.T. of the first kind, then (1-p) is the probability that an individual will experience no adverse effect. If, as a chronic "thought experiment" only (because it is physically impossible to conduct), an individual could be repeatedly retested using the same dose, adverse effects would be noted in some tests and not in others in a limiting ratio of p/(1-p): that is, the same individual would sometimes show and sometimes not show an adverse effect. This would not be a surprising result in that unpredictable variations in the receptivity of the individual can cause variations in the outcome to occur. If, on the other hand, the situation for an I.T. of the second kind is considered, then the measured p for a test with n animals is just a measure of the fraction of the test population having I.T.'s at or below the tested dose. In this case the measured p is not an individual probability, but rather the resultant of individual properties and it measures the probability that a randomly selected individual will be one that is adversely affected at or below the dose in question. In either case, dose-response will be observed but the existence of I.T.'s of either kind will not thereby be proved or disproved nor will it be possible to distinguish between them on the basis of chronic tests only. An extension of this logic leads to the conclusion that the latency argument as given above does not, in fact, prove the existence of either kind of I.T., however much it may at first appear to suggest it.

A parallel to the animal-testing outcomes described above may be useful to further clarify the reasoning involved, making use of the statistician's favorite example of sampling populations which consist of well-mixed white and black balls. Two cases are assumed corresponding to the two kinds of I.T.'s: (1) the first population consists of balls each of which has a probability equal to p of being black at any given time and equal to (1-p) of being white at any given time, and (2) the second population consists of a mixture of black balls which are always black and white balls which are always white and the fraction of black balls is p and of white balls is (1-p). If, under changes in some specified external influence, the probability of being black, p, changes in Case (1), and the fraction of black balls, p, changes in Case (2), then each case will exhibit the equivalent of dose-response; Case (1) corresponds to the thought-experiment described above (as for I.T.'s of the first kind) and Case (2) corresponds to the case of I.T.'s of the second kind.

In each case, a sample of n balls is taken and the proportion of black balls evident at the time of sampling is determined without further examination. In each case an estimate of p is obtained and, as larger or more numerous samples are taken, both estimates will converge on the single value of p corresponding to the level of the external influence (or "dose") imposed. Though as defined here the

true value of p is the same at a given level of external influence in Cases (1) and (2), the meaning of p, mechanistically, is not the same. Unlike the case of chronic animal experiments, the difference in meaning between the two cases is easily determined: a single black ball selected in Case (1) and repeatedly observed over time will show that it is sometimes black and sometimes white, whereas in Case (2) such repeated observations of a single black ball will show that it remains black. Such a simple determination of the existence of a difference in mechanism is not open to us in animal testing since a single animal cannot be tested repeatedly. Thus, it can be concluded that the normal, statistical, chronic animal tests, whether they exhibit dose-response or not, may suggest but cannot clearly demonstrate the existence of I.T.'s (or of A.P.T.'s) and biological information and studies of mechanism are needed to do so.[a]

Applying the above arguments to the latency-dose-response/ threshold argument given earlier, it is seen that the fact that an animal does not exhibit an adverse effect in its lifetime at a given dose says nothing about the existence or non-existence of either kind of I.T. for that animal since it can say nothing about mechanism: if the experiment could be repeated over and over again with the same animal, a response might occur, sometimes within the animal's lifetime and sometimes not [Case (1)], or else the same result might always be obtained [Case(2)].

There are thus two reasons why the usual chronic tests (and, for that matter, epidemiological studies) cannot be expected to clearly demonstrate the existence of I.T.'s (or, therefore, A.P.T.'s): (1) the fact that no real test can have sufficient statistical power to yield proof, and (2), the fact that such tests cannot distinguish between mechanisms but simply give an estimate of p without informing us as to what p means.

Although in principle the two cases can exhibit different variances even though each may have the same observable p for a given dose, enough measurements are not likely to be available in any real case to make such a means of distinguishing between the cases feasible, aside from the confounding effect of other experimental contributions to variance. If it should prove to be possible, at some point in time, to identify the "most sensitive members" of a population, then the difference between the two kinds of

---

(a) A different form of the above ball experiment may be clearer. Instead of sampling from two sets of balls, sample from two sets of coins of identical size, weight and shape, but which differ in that the first set consists of normal coins with one head and one tail, each, while the second set consists of an equal number of two-headed and two-tailed coins, well mixed. Random samples of each set, taken without turning coins over or other intimate examination, will give statistically indistinguishable results. However, the difference in the two sets of coins can be detected either by examining individual coins in each set, or by repeatedly tossing one coin from each set.

dose-response models would become important. Under such circumstances, any individual having a specific threshold, D_T, would be faced with very different personal understandings of risks, depending on whether the dose-response curve (and that individual's D_T) is of the first or second kind: if of the first kind, if a potential exposure is equivalent to $D>D_T$, there is an enhanced probability of cancer, whereas, if of the second kind, there is a 100% probability of cancer for the equivalent $D>D_T$. The personal decision as to whether the risk is worth the benefit, from the individual's perspective, will certainly be different in the two cases!

In spite of the fact that we are left in a quandary about the existence of I.T.'s because reasonable-seeming theoretical mechanisms can lead to models exhibiting I.T.'s and other reasonable-seeming theoretical mechanisms can lead to models which lack I.T.'s and because there are two reasons why the usual chronic studies cannot demonstrate the existence of I.T.'s, experimental studies of mechanism can provide the basis for concluding whether I.T.'s exist or not in a given case. For example, some agents can produce adverse effects by physical means such as insult to the lining of the bladder by solid particles. If the particles are bladder stones formed by precipitation of a material which requires an ingested agent for its formation, then for dose levels such that no stones are formed because of their

solubility in urine, the physical insult cannot occur. Such a grossly simple mechanism can result in an I.T.; more complex ones would require much ancillary experimental work for a reasonable demonstration. The CIIT has, in fact, recently found that hyperplasias are apparently formed in the bladder in tests with terephthalic acid and the stone-insult/solubility mechanism described above appears to be involved.

The absolute population threshold (A.P.T.) is that dose level below which no member of a population exhibits an adverse effect. For reasons given above, purely statistical determinations of dose-response cannot rigorously "prove" such a threshold exists. Even if I.T.'s are shown to exist for a given agent, individual variation can be so great that the requirement that no population member will suffer an adverse effect is not thereby proven. In the bladder studies mentioned above, the likelihood of the existence of an A.P.T. is high: even for extreme variations in urine composition or production rate, or for variations in stone composition, there may well be some level of dosage (and concentration) below which complete solubility is obtained. If the proof of the existence of I.T.'s is difficult, that for A.P.T.'s is even more so -- though not, in the ultimate, absolutely impossible if mechanism can be defined clearly. There has been in the literature some confusion, when

speaking of thresholds, between the I.T. and the A.P.T. It is important to bear the distinction in mind when debating the threshold issue.

The practical population threshold (P.P.T.) is a somewhat fuzzy concept which has not been defined. It could be very useful to define an acceptable one, however.

One view is that reasonable and expert persons, viewing a sufficient amount of dose-response data, and considering all other, relevant information, might conclude that an empirical curve such as Curve D should be drawn to intercept the abscissa, as shown in the Figure, and that T is a threshold, without serious risk of being wrong. While the words sound reasonable, the data for statistical reasons and even for unusually large numbers of animals will not permit one to show that the correct curve is D and not C (curves such as C are known). Further knowledge of biology and of the route(s) and mechanism(s) of carcinogenesis is needed, here, to assist in determining what the correct curve is likely to be. This type of threshold determination is one of consensus and depends on various considerations, including possible subjective ones. However, if data tend to show curves like the solid portions of C or D, the further major effort to study mechanism would be better justified than in the case of curves like the solid portions of A or B.

In the case of non-carcinogenic toxicological phenomena, a quantitative convention has been used to select what amounts to a practical threshold, the no-observed-effect-level (NOEL). Here, in its simplest form, the test dose at which no effect is found in a set of test animals (at a pre-selected confidence level) is divided by 100 (by 10 for intraspecies variation and by 10, again, for interspecies variation) to yield a kind of "safe" level or practical threshold. The difficulties with false negative results, repeated runs, etc., are obvious, but this device has served as a practical basis for action and has, no doubt, acted to contain risk. It has not been extended to carcinogenesis as a general rule.

Other ways to define a practical threshold depend on somehow determining if and where the dose-response curve changes slope abruptly upward (t, Curve C), a procedure of dubious utility and interpretation, or on selecting some level of risk below which (estimated by one of the various, conservative methods of extrapolation) risks are considered to be ignorable. Such risks have generally been set at what amounts to "near zero"; for example, 10^{-6} per lifetime. This, combined with very conservative extrapolation methods, causes such "thresholds" to be not so practical in that the allowable exposures so calculated may be practically unattainable yet at higher exposures the true risk may already be undetectable.

In the case of determining where a dose-response curve makes an abrupt change in slope, the "threshold" does not relate to the existence or non-existence of an adverse effect or even, as discussed further on, whether an effect is detectable or not detectable using normal statistical confidence limits in a population potentially at risk; it relates only to the acceleration of risk with dose (perhaps, to some change of mechanism) and the fact that a risk is below the knee need not mean that the risk is not significant: it all depends on how high t is in terms of p.

One further way to define a practical threshold would be to define it as that level of exposure which may not quite lead to a significant increase in cancer incidence, at a preselected confidence level, in the human population under consideration and, as applicable, for a given type or set of types of cancer. This definition, which would yield different numerical thresholds depending upon, among other things, the number potentially exposed in the specific population considered and how rare or common a given adverse effect is in the general population, also depends on being able to make and use a maximum likelihood estimator estimate of risk as a function of dose at the low probability levels of interest.

So far, no one method or convention for defining a P.P.T. has emerged that is well enough defined or justified as to make it possible for general adoption and use.

One final word on types of thresholds: beware of the false threshold. In the past some have noted that there are naturally occurring materials which, while they are carcinogens and exhibit dose-response, are needed in the body in small amounts for good health. These small amounts have been called "thresholds", when in fact they are no more than Nature's own solution to a risk-benefit problem.

Thresholds and Carcinogenic Safety

If an A.P.T. or an I.T. can be shown to exist and the effort to demonstrate its existence produces either a better function for extrapolation or a dose-range in which the A.P.T. (or the I.T.'s) might lie, or even both at once, the advantage in cutting down the uncertainty of risk assessment is obvious and very great. One can then more easily make a reasonable, maximum likelihood extrapolation and say something meaningful about the size of the uncertainty involved.

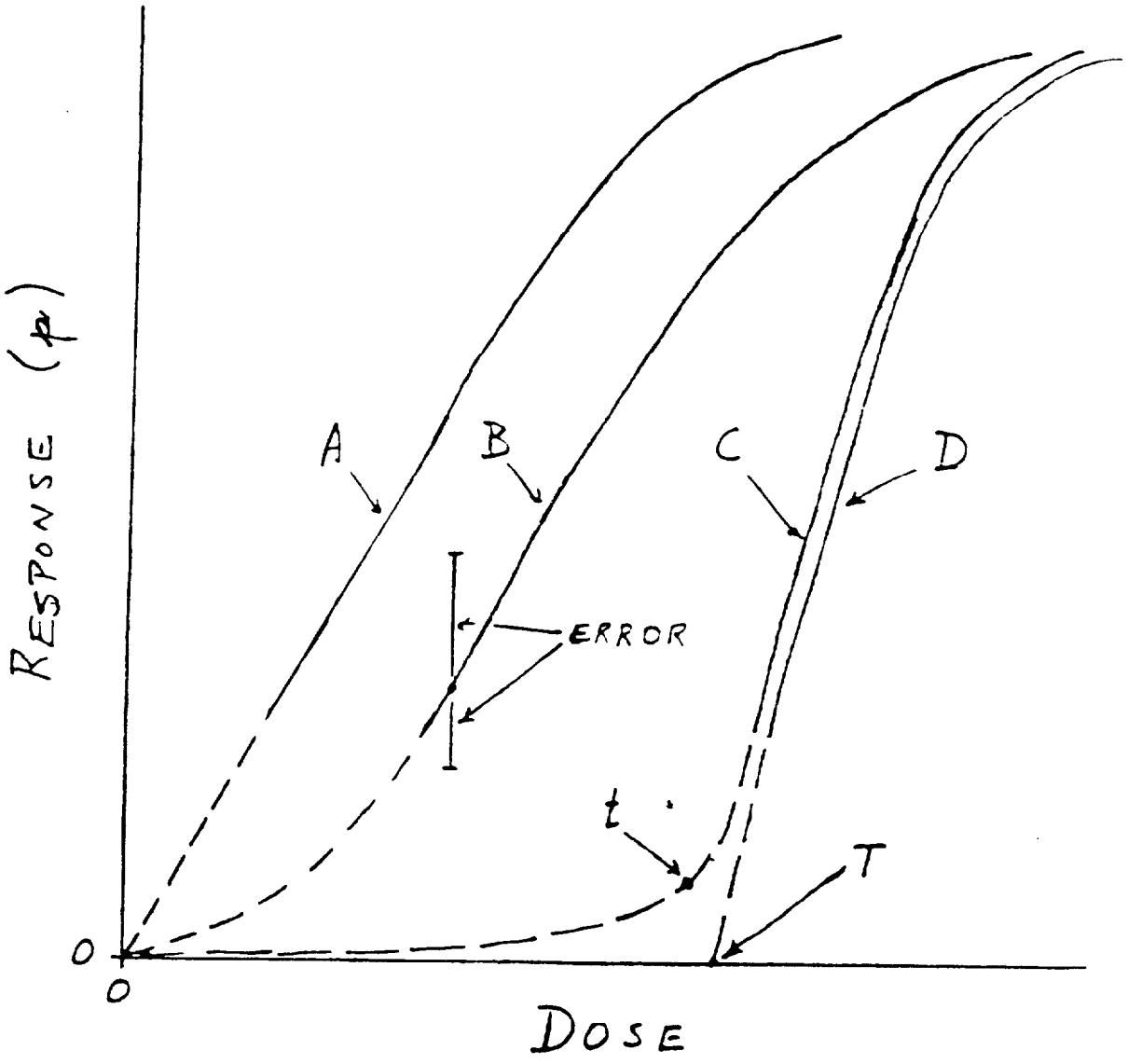
Knowing that there is an A.P.T. or a set of I.T.'s and where they might lie does not necessarily make it possible to find a truly "safe" dose: if the A.P.T. or I.T.'s are so low as to pose problems of attainment of the resultant low exposures in practice, especially recognizing that a safety

factor would prudently be applied to determine an "allowable" level, the knowledge would have the advantages listed above but we would still have to go through the exercise of determining what the acceptable level of risk is: the problem would be like the "no-threshold-by-policy" case, but using an extrapolation function more appropriate to the task. In the case of an A.P.T. or a set of I.T.'s which, after application of a safety factor, lead to a readily attainable exposure, the advantage in knowing them can hardly be overstated.

The investigations required to determine the existence and, if applicable, the likely location of an A.P.T. or set of I.T.'s are not routine in any sense and will be likely to take different turns in every case. Indeed, much effort may even lead to answers of equivocal value in various cases. This being the case, it would generally be prudent to give such an effort high priority only when confronted by data exhibiting steep slopes as in the solid portions of Curves C and D. For curves like A and B it would be less likely to be of profound benefit, though each case would require a separate decision involving such factors as the value of the result. Conducting the experimental effort in this latter case would generally but not always tend to be given a lower priority.

The determination of an A.P.T. will generally be arduous, elusive (because of the definition of the A.P.T.), uncertain as to process, and uncertain as to outcome -- though the data acquired along the way will very likely have utility in any case. The time necessary to make and interpret the experiments is also unpredictable, and if a real hazard is thought to exist, action of some sort may be necessary long before the determinations proceed very far. The use of a set of reasonably likely fitting functions for extrapolation, which give similarly good fits to the experimental dose-response data to estimate the range in which it is thought the risks might lie at different dose levels together with a P.P.T. of the "not-quite-significant-effect-for-the-population-considered" type, with a safety factor chosen on a case-by-case basis and with the further, usual considerations of cost, feasibility and the reality of potential health improvements to be achieved may prove to be a practical approach, achievable without major delay. The studies to determine, as nearly as possible, the existences of I.T.'s and A.P.T.'s, unless the mechanism is an unusually straightforward and simple one as in the case of terephthalic acid cited above, would more likely be conducted after at least some control actions have occurred and then only in the case of some very important material the wider or easier use of which offers real advantages.

Even then, such studies are likely to be made only if there is some hope of success based upon data already available as described above.



VARIOUS DOSE-RESPONSE FUNCTIONS

PFD 8/20/80