Environmental Health Perspectives
Vol. 26, pp. 37-42, 1978

Epistemology of Screening for Behavioral Toxicity

by P. B. Dews*

A method is described for the assay of the behavioral effects of volatile solvents on mice and illustrated with pilot results on trichlorethylene. A dose-effect curve has been determined for the effects on schedule controlled responding and compared with the dose-lethality curve and the TLV for man. The OR_{50} for behavioral effects was $^1/_5$ of the LD_{50} and 50 times the TLV for long-term exposure of man.

An analysis of the errors involved in determination of effects on whole animals leads to the conclusion that subtle effects, representing a few per cent change, will not be detectable in routine screening. It is suggested nevertheless that information on the midrange, knowable, part of the dose-effect curve may prove useful in predicting safe levels for man.

What methods have been shown to predict when prolonged exposure to a low level of an agent will lead to subtle and delayed behavioral effects in man? None. The present communication describes methods that on the basis of their other uses and of the results of pilot experiments should be assessed as one means of providing information helpful in the rational development of limits for human exposure. Inevitably, the epistemology of testing will be discussed and, finally, general recommendations will be made.

For chronic toxicity, the small, cheap, plentiful mouse has been used extensively, so familiar behavioral techniques have been applied to the mouse. The techniques involve: an objectively recorded response and automatic programming of sessions; relatively long periods of observation (ca. ½ hr) allowing time for the response to occur many times, giving a good sample of behavior for quantitation; standardized training leading to steady-state responding from session to session over long periods of time (months or years); applicability to all mice of all strains; no selection of individuals, no discards.

Apparatus

The apparatus is a modification of an apparatus already described for use in behavioral pharmacology (*1*). It consists of a small cage with a blind corridor leading off one end. In the floor of the corridor is a hole to which a dipper containing evaporated milk, undiluted from can, can be brought. Above the dipper hole is a light beam shining across the corridor onto a photocell. When the beam is interrupted by the mouse, there is an audible click from operation of a feed-back relay, and a response is recorded.

Schedule

A mult FR FI schedule is imposed. In the presence of a stimulus such as brief bursts of white noise at 6 Hz, 30 responses lead to a dipper of milk for 10 sec (FR 30), plenty of time for the mouse to consume the 0.05 ml of milk and to lick the dipper clean. After 30 sec of quiet darkness (TO) another stimulus appears, say a light, for 300 sec, then food again when the beam is broken (FI 300 sec), then 30 sec TO, then FR 30, and so on. A standard session is five sequences of TO—FR 30—TO—FI 300 sec; then a 600 sec pause, and then five more TO—FR 30—TO—FI 300 sec. In acute experiments, drugs or toxins are given at the beginning of the 600 sec pause. In chronic experiments it may be enough to conduct one sequence of 5 per day, requiring a little over one-half hour. Performance is monitored continuously with a cumulative recorder and additionally the rates of responding in the FR and FI components are separately recorded.

* Laboratory of Psychobiology, Harvard Medical School, Boston, Massachusetts 02115.

Procedure

A standardized training sequence should have two main features: it is unequivocally specified, leaving nothing to discretion, and it should lead effectively to the training of every individual subjected to it.

An effective standardized training sequence for mice that has proved similarly effective with C57Bl and C-D 1 mice is the following. (1) The mice are deprived of all food for 60 hr (in practice, Friday PM to Monday AM). (2) The mouse is put in the apparatus, subject to the following schedule. The sequence of stimuli is as in the definitive schedule described above (6 Hz white noise, TO, light, TO, and so on) but the first time beam is interrupted in presence of either white noise bursts or light leads to food. (In other words, the schedule is mult FR 1, TO, FI 0 sec, TO). If a response is not made within 30 sec of the start of white noise bursts, the sequence shifts to the next component, TO, then FI 0 sec. If a response is not made within 330 sec of light period, the sequence shifts to next component, TO, then back to FR 1 and so on. A session comprises 15 FR 1 and 15 FI 0 sec components. (3) The above procedure is repeated in subsequent daily sessions until food is presented 30 times, that is, subject never fails to respond for long enough in either component for sequence to change spontaneously. One or two sessions ordinarily suffice. (4) The schedule is changed to mult FR 1, TO, FI 300 sec, TO. The FR component cannot last longer than 30 sec and the FI component cannot last longer than 330 sec. The sessions are repeated until at least 27 food presentations are made. (5) The FR parameter is increased through following steps: 2, 3, 5, 10, 20, to 30 in subsequent sessions, always provided that 27 food presentations occur. If less than 27 food presentations occur, the FR parameter is not increased for next session. (6) When the FR parameter has reached 30, the final program, described above, is imposed in two half-sessions each of 5 FR 30 and 5 FI 300 sec.

The training sequence has brought every mouse on whom it has been imposed, and who has survived enough sessions, to differential responding under FR 30 and FI 300 sec, almost invariably in 12 to 20 sessions. Studies in behavioral pharmacology have then been conducted.

Pilot Experiments in Acute Behavioral Toxicology

To appraise the applicability of the procedures to toxicology, pilot experiments have been made with trichlorethylene (TCE). The apparatus was so modified that all exposed surfaces, except that of the milk, was aluminum, and the apparatus enclosed in an aluminum-lined chamber (a picnic ice box, in fact) of about 40 l. capacity. There was a fan in the chamber, and a small, closable aperture introduced in the roof for addition of measured amounts of TCE. A mouse was introduced into the chamber, subjected to five sequences of FR 30, TO 30 sec, FI 300 sec, TO 30 sec. TCE was then introduced into the chamber, the fan run 10 min, then the second five sequences imposed. Simple calculations indicate that a mouse can cause only infinitesimal changes in p_{O_2} and p_{CO_2} in the chamber in the course of an hour or so, and the calculations have been vindicated by showing that mice survice 24 hr continuously in the hermetically sealed chamber without observable effects. The concentrations of TCE in the chamber are calculated from the volume of the chamber and the amounts of TCE added, assuming no loss, so they represent upper bounds of concentrations. It was possible to have access briefly to a GC analyzer that indicated that, for toluene, the actual concentrations at the end of half an hour approximated the theoretical values, the loss being in the vicinity of 20 or 30%.

Results

Rates of responding under FI 300 sec averaged a little more than 0.5 responses/sec under control conditions. The rate was reduced almost by half in the second five sequences by the 10-min exposure to the fan; the rate was reduced a further 50% by about 23 g/m³ of TCE; let us take this figure as a first approximation to the OR_{30} (Fig. 1), where OR_{50} denotes the dose reducing the output of responding to 0.5 of its control level.

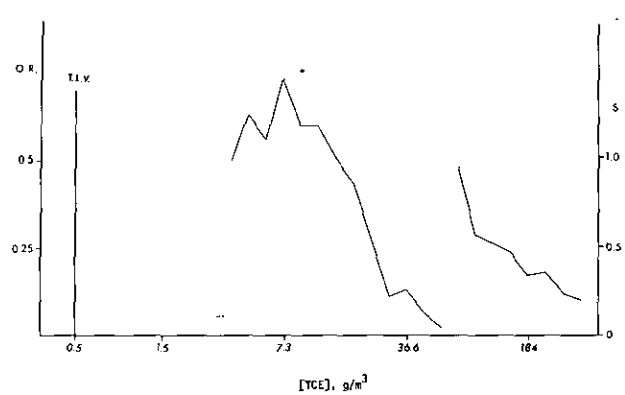

FIGURE 1. Effects of TCE: (right curve) lethality (scale of proportion of survivors on right ordinate); and (left curve) suppression of responding under FI 300 sec (scale OR on left ordinate). Also shown is TLV for human workers.

To put this value in perspective, the LD_{50} for TCE was determined for mice exposed for 40 min in the same chamber. The LD_{50} was a little over 100 g/m^3 TCE, which is about five times the OR_{50} for the behavioral effects. The slope of the dose-effect curve is about 6 logits per $\log_{10}$ unit change on concentration and that of the dose-lethality curve is about 4 logits per $\log_{10}$ unit change in concentration.

The TLV of TCE for chronic exposure to human subjects is 0.5 g/m^3, about $^1/_{50}$ of the OR_{50} for mice.

The relative concentrations for LD_{50}, OR_{50}, and TLV seem not unreasonable. General anesthetics produce anesthesia at about half the lethal concentration, and it seems reasonable to expect a substantial behavioral effect (an OR_{50} effect) at rather less than half the anesthetic concentration, so the TCE results are similar to the known effects of the volatile organic solvents that are used in anesthesia. It also seems reasonable that chronic exposure to man should not exceed one-fiftieth the concentration profoundly affecting mice acutely.

The dose-effect curve for TCE was determined on six mice at 12 dose levels spaced 1 decibel (0.1 $\log_{10}$ units or 1.26-fold) apart. Close spacing was chosen to obtain information on the precision of the assay. It is clear from Figure 1 that in the steep part of the dose–effect curve it was easy to detect differences in dose of as little as 1 decibel.

Each day, the first half-session is under control conditions, and the performance gives evidence on whether a mouse has recovered from the effects of treatments on previous days. With such a control, it is possible to give exposure to the agent three or even four times per 5-day week, and so it is theoretically possible to obtain a dose–effect curve like the one shown in about four working weeks. In routine operation in the study of new agents, a wider spread between dose levels would suffice, 5 or 10 decibels (3-fold or 10-fold) and given the possibility of a sequential design, where additional dose levels are determined in the light of the effect of levels already studied, fewer dose levels would be necessary. With a platoon of trained mice, the establishment of dose–effect curves for the acute behavioral effects of agents at the rate of one agent per month per apparatus seems attainable. One technician could tend five or six apparatuses, so reasonably efficient generation of information would be possible.

Error Estimates

The initial control half-session each day provides much information on day-to-day, long-term, and mouse-to-mouse variability. Some 12 mice were studied over 30 or more sessions. The mean rate of responding in FI 300 sec over the 30 control half-sessions was 0.58 responses/sec, with a range over the mice from 0.35 to 0.77, a standard deviation of 0.143, and so a coefficient of variation of 0.25. Assuming normal distribution of error, simple large sample statistics suggests that a 20% effect of an agent could be detected with six mice and a 10% effect with as few as 25 mice for $p \geq 0.05$. Remember, however, that the variances are based on the mean of 30 determinations in each mouse, so that although only six mice may be needed, 180 observations are involved for 20% difference and 750 observations for 10% difference. The numbers are sobering but not numbing. For a 1% decrement, however, which is more than one would care to see in a human subject on a chronic basis, 75,000 observations would be necessary. Also, the numbers apply to the detection of an acute effect. Detection of an effect that develops only after 30 or 100 or 300 days of exposure requires a tremendous upscaling of exposure and holding facilities.

Actually, the real situation is both better and yet impossible. The above calculations were made on the basis that the standard error of a population of means of sample size n drawn from a parent population goes down as n is increased, such that $SE = SD/\sqrt{n}$. In the pilot experiments in the mice, 12 mice were studied over 30 sessions, giving 360 observations. For each mouse, the standard deviation was estimated from three consecutive series, each of 10 consecutive observations. The standard deviation of the means of 10 consecutive observations was also estimated directly from the three means for each mouse. The former standard deviation should be $\sqrt{10} = 3.16$ times larger than the latter; actually, it averaged only 2.38 times larger over the 12 mice. Hence, fewer than 30 observations will cause less of an increase in error than would have been expected. That is the better aspect of the real situation.

The impossible aspect of the situation is that the failure of the estimated standard error to be reduced by $\sqrt{n}$ is a clear indication that there are errors additional to sampling errors. If a "real" (i.e., nonsampling) error has an average size of ϵ and a probability p of occurring and always has the same sign, as well it might, then the total error in n observations would be expected to be $np\epsilon$ and the average error $p\epsilon$. Note that n does not appear in the latter expression; the average error, $p\epsilon$, does not go down with sample size. The mathematics are inelegant but the message clear. It is an ineluctable feature of real life experiments that real errors have real effects on results. Actually, the difference between 2.38 and 3.16, above, is unusually small. Many years ago it was reported that the error variance for replication of experiments was two to three times the variance

estimated from variability within an experiment (2). The present experiments on mice were conducted by one person in a single laboratory and single apparatus, and the variances refer to replications in individual subjects. It is likely that under less rigid circumstances, as must obtain in extensive testing, the real variance will be considerably greater. Schneiderman, Mantel, and Brown (3) have discussed some of the reasons that real errors cannot be indefinitely reduced by increasing n; basically, accidents must happen and mistakes must be made, both in conduct of experiments and chronicling of results. When real data are used to estimate real errors, the errors turn out, in biomedical work at least, to be far higher than anybody's worst fears. As indicated, real error variances two or three times sampling variances were found. In designing experiments, especially long-term experiments, I suggest that to assume a real error variance only as large as sampling variance is as optimistic as we dare be. The results in the literature clearly place the burden of proof on those claiming lower errors.

Implication

The implication, quite simply, is that it is impossible to perform an experiment that will measure a small biological effect of the type of low probability lethality or slight biochemical, physiological, or behavioral effect that depends on an assessment of the whole subject. The estimated coefficient of variation in the mouse pilot experiments was 0.25. Assume half of this is nonrandom and does not decrease with n. A 30% decrement should be detectable by as few as 25 mice, but a less than 25% decrement could never be detected.

It is my hunch that, in practice, for most whole animal experiments, if an effect cannot be detected with an n of 30, then it cannot be detected. Many years ago, Burn wrote (4): ". . . in practice groups of not less than 30 should be used. With smaller numbers the error is too great; larger numbers are impracticable." Burn was no statistician but he was unrivalled as an assayist; and he knew what was practicable. Of course, the rule of 30 will apply to only a limited type of experiment and there will be exceptions, but it is useful in that people claiming exemption for their experiments will be required to prove it.

We need to predict when subtle and delayed behavioral effects will occur from prolonged exposure to a low level of an agent. By subtle is meant, say, a few per cent decrement in performance. The foregoing discussion on errors suggests that we cannot determine what levels of an agent cause only a subtle effect of a few per cent decrement by exposing more and more subjects to lower and lower concentrations to determine directly points on the extremely low portion of the dose–effect curve. Further, mathematical extrapolation is generally agreed to be impossible: the different models that have to be assumed to make extrapolation possible (probit, logit, etc.) are impossible to differentiate in the middle range of the dose effect curve, where experimental determinations can be made, yet lead to different estimates of, e.g., OR 10^{-6}.

Is prediction of low-level effects, therefore, impossible? Perhaps, but not necessarily. We have learned a great deal about matters not directly determinable, as in astronomy. What information can we collect from the realm of the possible, that, as earthbound chemistry, physics and spectroscopy lead to firm inferences on the composition of the sun, will let us peer confidently into the far reaches of the dose–effect curve?

Dose–effect curves in animal subjects over the range of effect from 0.2 or even 0.1, to 0.8 or even 0.9, can be determined, along with their slopes and variances. We can determine lethality curves. We can determine dose–effect curves over the mid-range, even for chronic experiments. We have a large amount of clinical information on effects in man, and some good epidemiological information. We could make monitors that would tell us about exposure both from time to time and as an integrated average. Can we attain some predictive capacity from correlations? Is there, for a particular class of compounds, a consistent relation between position and slopes of dose–effect and lethality curves and the TLV (which we can look up in a book)? What would figures like Figure 1 look like for other agents? I realize that many subjective judgements go into the establishment of a TLV, but would it not be interesting if for many agents there were reasonable relations between animal subject OR_{50}, LD_{50}, and the published TLV? Should not anomalies then be re-examined? I submit such experiments can and should be done before we abandon the hope of prediction and before regulations are promulgated.

A Last Suggestion

A frequent debate in toxicology that will soon involve behavioral toxicology, is the use of pure strain mice versus heterogeneous populations. The basic argument is that since human populations are heterogeneous, it will be easier to predict human effects from heterogeneous mouse populations despite the fully recognized much greater variability of the heterogeneous mice. Previous discussion shows why the disadvantage of greater variability cannot

be overcome simply by increasing n.

Suppose we have two strains of mice that generate dose effect curves as shown in Figure 2. Note the two curves have the same slope, 4 logits/$\log_{10}$ unit, which is the slope estimated for the lethal effect of TCE in mice, but the OR_{50} values of the two strains differ by a factor of 10. If we had a heterogeneous population consisting of 50% A and 50% B, the dose-effect curve for the heterogeneous population would be the middle line. It is curvilinear, although that is not likely to be detectable in real experiments. The estimate of the slope from +2 to −2 logits (0.12 to 0.88) is 2.44 and from +3 to −3 logits (0.047 to 0.953) is 2.75. The true slope at doses below 1.0 is, however, indistinguishable from

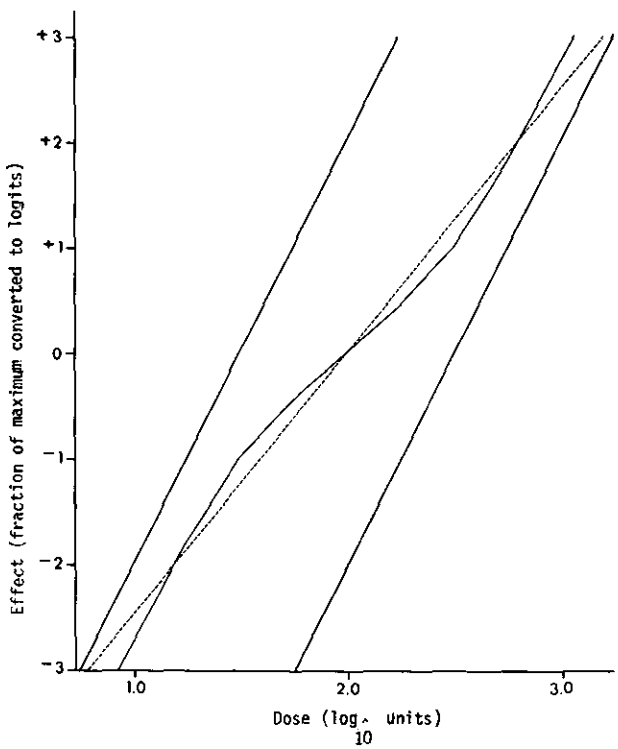

FIGURE 2. Theoretical curves illustrating what may be result of mixing mice of different strains of different susceptibility. The effect is shown as fraction of maximum converted to logits, i.e., if the proportion of maximum effect is p, the logit is ln $[p/(1-p)]$. The two parallel straight lines are theoretical dose-effect curves for two strains of mice, A on left and B on right. Both lines have slope of 4 logits/$\log_{10}$ unit but A is 1 $\log_{10}$ unit more sensitive than B. The center curvilinear line C shows the dose-effect curve for a population consisting of 50% of strain A and 50% of strain B. The dotted line (D) shows the slope of the dose effect curve for the mixed population that would be estimated from observations from doses with effects of from −2 logits to +2 logits (0.12 to 0.88 of maximum). Note that for doses below 1.0 the curve for the mixed population has become virtually parallel with the curves for the pure strains.

4. This is because as the dose is reduced, the contribution to the effect of members of strain B becomes vanishingly small. The estimate of the effect of small doses is much poorer from the combined strain mice, estimated over the practical range, than from strain A. With a really heterogeneous population, and not just a mixture of two pure strains, we may anticipate that, as the dose is lowered, only individuals with a more and more restricted collection of the relevant genes are affected; that is, the dose-effect curve more and more closely resembles that of a pure strain. It may therefore be practical to estimate the effect of low doses on a heterogeneous population as follows: determine the slope of the dose-effect curve on a pure strain; determine the OR_{50} and the standard deviation of the heterogeneous mice on the heterogeneous population; draw a line with the determined slope but, say, three standard deviations to the left of the OR_{50} point for the heterogeneous population; then extrapolate cautiously. Let us try for 0.05 and 0.01 before we head for 10^{-6} and 10^{-9}; let us look to know something about the sun before aspiring to weigh distant galaxies.

Conclusions

The assessment of the subtle and delayed behavioral effects of prolonged exposure to low levels of agents is urgently needed. It is suggested that it will be impossible to study enough animal subjects, long enough and with low enough experimental error to determine directly effects of agents causing changes of only a few per cent. To determine the possibility of indirectly determining levels the following information should be generated and assembled for a variety of extensively used agents: the dose-effect curve for behavioral effects in a number of both pure strain and wild type mice and the dose-lethality curves for the same mice.

Slopes and OR_{50} and LD_{50} values should be calculated.

These figures should be assessed in the light of all available information on man. Efforts should be made to improve information on actual exposure in humans, and to find simple objective means of detecting behavioral deficiencies in man. I realize these last two are monumental tasks, but work on them will be cumulative and give hope of providing eventually a basis for predictive testing. No methods of predicting when prolonged exposure to a low level of an agent will lead to subtle and delayed behavioral effects in man have been validated. Until they have, granting agencies should be strongly urged to make funds available for studies that will help the development and validation of

methods and regulating agencies should be strongly urged not to require the use of unvalidated methods. Rules of testing promulgated today have an infinitesimal chance of specifying methods that are optimal, and the required use of less good methods will consume the resources that should be going into the development of better ones.

The experiments on the behavioral effects of TCE in mice and the preparation of this paper were supported by the Stanley Cobb Fund of Harvard University.

REFERENCES

1. Wenger, G. R., and Dews, P. B. The effects of phencyclidine, ketamine, *d*-amphetamine and pentobarbital on schedule-controlled behavior in the mouse. J. Pharmacol. Exptl. Therap. 196: 616 (1976).
2. Dews, P. B., and Berkson, J. On the error of bio-assay with quantal response. In: Statistics and Mathematics in Biology, O. Kempthorne et al., Eds., The Iowa State College Press, Ames, Iowa, 1954, pp. 361-376.
3. Schneiderman, M. A., Mantel, N., and Brown, C. C. From mouse to man—or how to get from the laboratory to Park Avenue and 59th Street. Ann. N. Y. Acad. Sci. 246: 237 (1975).
4. Burn, J. H. Biological Standardization, Oxford Univ. Press, London, 1937, p. 35.