

March 31, 2000

**UP AND DOWN PROCEDURE:
IS THERE NEED FOR FURTHER
COMPUTER SIMULATIONS AND *IN VIVO* VALIDATION?**

BACKGROUND

Acute Oral Toxicity Testing

The acute oral toxicity test seeks to estimate the dose at which 50% of the organisms in a defined population will die (LD50) after exposure to a test material. The statistical basis for the classic study design was first described in the 1920s and remained in use until current times. In this test, groups of animals were administered varying doses of test material, and a dosed animal either lived or died. As the dose in an acute toxicity test is increased, the probability that a given animal dies increases. These results established a relationship between dose and response. Responses in an acute toxicity study can be characterized by a mean (the LD50) and variance (or slope) of the dose-response curve.

Over the years many attempts have been made to expand test outputs and to adjust statistical sampling so as to minimize the number of animals used and decrease their pain and suffering. These changes in sampling technique do not involve any change in the actual treatment of the animals or the lethal endpoint of the test. Over the years, the classic LD50 protocol has been modified to reduce the number of animals from scores of animals to 15 to 30 per study. Other modifications include such things as:

1. The dose is usually administered by oral gavage to fasted young adult animals.
2. Animals are observed periodically during the first 24 hours with special attention given to the first four hours, then at least once a day for 14 days or until they die or recover.
3. Clinical signs including their nature, severity, time of onset and to recovery are recorded at observation times.
4. Body weights are determined before treatment, weekly thereafter and at death.
5. All animals that survive are sacrificed at 14 days.
6. Gross necropsies are done on all animals in the study; histopathology of lesions and clinical chemistries may be included.

Response Variability

Variations in results from a study of a given chemical can be divided into many different components:

1. animal age, sex, estrus cycle, strain and species
2. among animals in a study

3. among groups of animals in a study
4. studies at the same or different times within a laboratory
5. studies conducted in different laboratories.

It is recognized that as long as the animals in a test are individually housed, the animal to animal variability and variation with age, sex, strain and species will not change with the sampling procedure, i.e. for protocols with sequential vs. simultaneous dosing. It is important that adequate population variability be built into the computer simulations and enough is known about the endpoint to be able to write a computer program that can accurately predict experimental results.

Computer Simulation as an Aid in Test Design

An experimenter wants to use sampling designs with small numbers of animals which adequately estimate the mean and variance of the entire population. When both the mean and variance of the population are known, it is possible using a computer to run the specified test hundreds or thousands of times by generating random sequences of responses. Thus, the computer simulates overall results by repeatedly taking small samples from a much larger population. Simulations provide a way to select among designs those with the greatest accuracy in estimating the mean and variance (or standard deviation) of the population. No level of in vivo testing could ever generate the number of runs that are possible using simulation.

In Life Testing

Certain aspects of test designs may not be totally addressed by computer simulations. In going from theory to practice, there are other considerations. For instance, for each design, has the protocol been ably articulated so that laboratories can consistently carry out the study and accurately assess study outcomes? Without some laboratory experience it is not possible to unequivocally assert that the method can be appropriately utilized. Generally, some laboratory information is needed to confirm that a new test method performs in the way hypothesized against a "gold standard" method. Likewise, across acute toxicity designs, there is similar variability within and among laboratories. The same is the case for variability within a laboratory over time. However, if the test method is the same across various toxicity test designs, there should be similar variability within and among laboratories. The same is the case for variability within a laboratory over time.

UP AND DOWN PROCEDURE (UDP)

Significant work has been performed on the UDP. Theoretical studies have demonstrated the characteristics of the method and indicated that the procedure and its modifications are the most efficient means of deriving an estimate of the median effective dose per expenditure of test animals (Brownlee et al., 1953; Wetherill et al., 1966; Dixon, 1965; Hsi, 1969; Little, 1974a,b). Practical determinations of acute toxicity bear this out, where savings in animals in comparison to the classical test and the FDP can be significant; the UDP and the acute toxic class method appear to use quite comparable numbers of animals (Bonnyns et al., 1988; Brownlee et al., 1953; Bruce, 1985, 1987; Yam et al., 1991; Schleder et al., 1994; Lipnick et al., 1995).

Data from 35 published test materials have been summarized which compare the UDP, which were assumed to have a sigma of 1.2 which is representative of many consumer chemicals, with the classic or other acute oral toxicity designs (Lipnick et al., 1995). This number of compounds for validation studies is similar to that run for some other acute toxicity and eye irritation validation studies. The results of these studies showed the UDP design was most often able to predict the LD50 determined by the classical LD50 test. The method was accepted as an American Standard Test Method and by OECD (1997) without further testing and validation (U.S. EPA, 1995)

However, there have been indications that all OECD acute toxicity methods, including the UDP, would not provide necessary information about all types of compounds and mixtures. During an evaluation in spring, 1999 of the four acute oral toxicity designs already accepted by OECD, all were shown by simulation techniques to have poor ability to estimate the LD50 of the underlying population when the slope of the dose response curve is shallow and the starting doses for the tests were far from the actual LD50.

Subsequently, the U.S. was asked to determine if improvements in the sampling technique could be made that would improve the ability of the UDP to estimate the LD50 of the underlying population. Modifications have been developed which adjust the design of the UDP regarding the spacing of doses, add rules for the cessation of animal testing and formulate a more efficient use of animals in a limit dose test. In addition, proposals for generation of dose response slope determination have been developed. It is recognized that the new proposed UDP is more complicated than that in the current OECD guideline.

Significant numbers of simulations have been performed to justify the new designs of the UDP. However, no *in vivo* testing has been performed to illustrate the applicability of the designs. Likewise, there have not been any comparisons of the new UDP and the classic LD50 design. Some believe that the extensive simulations provide data representative of the population which an animal experiment replicated few times will not provide. Others believe that it is critical to observe that the method can be used successfully in a laboratory, considering the complexity of the proposed method and the fact that the results obtained reflect computer simulations. The Pesticide Program of EPA has a substantial database of classic acute toxicity test results, some with repeat tests done by independent laboratories, that could be used as a comparison for actual *in vivo* UDP.

QUESTIONS FOR THE PEER PANEL

It is recognized that many further studies on the performance of the proposed UDP procedures could be undertaken. Some of them might include such things as:

1. ability to transfer the test method among laboratories
2. actual performance of the method with chemicals of steep and shallow slopes
3. actual performance of the method with chemicals from different toxicity categories

4. practicality of the UDP or other sequential dosing methods for chemicals with somewhat delayed deaths ?
5. impact on test results of changing animal age and weight which could occur for chemicals with delayed toxicities or shallow slopes?
6. outliers. Simulations can show the impact of many outlier responses. However, when one animal is tested at each dose, how would outlier responses in the laboratory be identified by the investigator or the regulatory agency?
7. inability of small sample size designs being able to identify the breadth and severity of toxic signs
8. comparison of the ability of the new UDP test and the classic design to predict chemical hazard classification
9. real life test variability, in comparison to that predicted from simulations
10. determine that the relevant ICCVAM criteria for validation have been reached
11. get information on chemical mixtures as compared to single substances.

Recognizing that any number of these areas could be investigated with further simulations or in vivo tests, the peer panel is asked to provide comment and recommendation on the following questions.

1. Are the simulations that have been performed appropriate for demonstrating the operating characteristics of the modified UDP? Are there further simulations that would be helpful in evaluating the strengths and weaknesses of the method?
2. Are there in vivo tests that would aid in the determination of the usefulness of the proposed test procedures?
3. If there are further simulations that would be helpful in ascertaining the usefulness of the test proposals, provide guidance as to the priority that they should receive, given that resources for further investigations are limited.
4. Is a limited in-vivo validation necessary to (a) determine practical applicability of this complex method in a contract laboratory, including influence of variables such as changes in animal age/weight in the course of the test or effect of changing animal batches to stay within age/weight range; (b) determine the performance of the method relative to confidence intervals of simulations and © compare in-vivo results with LD50 values available from existing data bases.

REFERENCES

ASTM 1987 (American Society for Testing and Materials) Standard test method for estimating acute oral toxicity in rats. Designation: E 1163-87. Philadelphia: American Society for Testing and Materials.

Blick, D.W., Murphy, M.R., Weathersby, F.R., Brown, G.C., Yochmowitz, M.G., Fanton, J.W., & Harris, R.K. 1987a Primate equilibrium performance following soman exposure: Effects of repeated daily exposure to low soman doses. Report USAFSAM-TR-87-19. Brooks Air Force Base, TX: USAF School of Aerospace Medicine. 18 pp.

Blick, D.W., Murphy, M.R., Brown, G.C., Yochmowitz, M.G., & Farrer, D.N. 1987b Effects of carbamate pretreatment and oxime therapy on soman-induced performance decrements and blood cholinesterase activity in primates. Report USAFSAM-TR-87-23. Brooks Air Force Base, TX: USAF School of Aerospace Medicine. 12 pp.

Blick, D.W., Murphy, M.R., Brown, G.C. & Yochmowitz, M.G. 1987c Primate equilibrium performance following soman exposure: Effects of repeated acute exposure with atropine therapy. Report USAFSAM-TR-87-43. Brooks Air Force Base, TX: USAF School of Aerospace Medicine. 11 pp.

Bonnyns, E., Delcour, M.P. & Vral, A. 1988 Up-and-down method as an alternative to the EC-method for acute toxicity testing. Brussels: Institute of Hygiene and Epidemiology, Ministry of Public Health and the Environment. IHE project no. 2153/88/11. 33 pp.

Brownlee, K.A., Hodges, J.L. & Rosenblatt, M. 1953 The up-and-down method with small samples. *J. Amer. Statist. Assn.* 48: 262-277. •6

Bruce, R.D. 1985 An up-and-down procedure for acute toxicity testing. *Fundam. Appl. Toxicol.* 5: 151-157.

Bruce, R.D. 1987 A confirmatory study for the up-and-down method for acute toxicity testing. *Fundam. App. Toxicol.* 8: 97-100.

Cordts, R.E. & Yochmowitz, M.G. 1983 Antiemetic studies both pre and post exposure: Preliminary findings. Report USAFSAM-TR-83-23. Brooks Air Force Base, TX: USAF School of Aerospace Medicine. 9 pp.

Dixon, W.J. 1965 The up-and-down method for small samples. *J. Amer. Statist. Assoc.* 60: 967-978. Hsi, B.P. 1969 The multiple sample up-and-down method in bioassay. *J. Amer. Statist. Assoc.* 64: 147-162.

ICCVAM. 1997 Validation and regulatory acceptance of toxicological test methods. A report of the ad hoc Interagency Coordinating Committee on the Validation of Alternative Methods. NIH publication no: 97-3981. National Institute of Environmental Health Sciences: Research Triangle Park, NC.

- Klaasen, C.D. & Plaa, G.L. 1967 Relative effects of various chlorinated hydrocarbons on liver and kidney function in dogs. *Toxicol. Appl. Pharmacol.* 10: 119-131.
- Lipnick, R.L., Cotruvo, J.A., Hill, R.N., Bruce, R.D., Stitzel, K.A., Walker, A.P., Chu, I., Goddard, M., Segal, L., Springer, J.A. & Myers, R.C. 1995 Comparison of the up-and-down, conventional LD50, and fixed-dose acute toxicity procedures. *Fd. Chem. Toxicol.* 33: 223-231.
- Little, R.E. 1974a A mean square error comparison of certain median response estimates for the up-and-down method with small samples. *J. Amer. Statist. Assoc.* 69: 202-206.
- Little, R.E. 1974b The up-and-down method for small samples with extreme value response distributions. *J. Amer. Statist. Assoc.* 69: 803-806.
- Meyer, J.H., Elashoff, J., Porter-Fink, V., Dressman, J. & Amidon, G.L. 1988 Human postprandial gastric emptying of 1-3 millimeter spheres. *Gastroenterology.* 94: 1315-1325.
- OECD. 1997 OECD guideline for the testing of chemicals. Acute oral toxicity: Up-and-down procedure. OECD guideline 425. Organization of Economic Cooperation and Development: Paris.
- Schlede, E., Diener, W., Mischke, U. & Kayser, D. 1994 OECD expert meeting: Acute toxic class method. January 26-28, 1994, Berlin, Germany.
- U.S. EPA 1995 Rationale for the up and down procedure. Submission to OECD concerning the acceptance process for the method. (Included in ICCVAM review package)
- Wetherill, G.B., Chen, H. & Vasudeva, R.B. 1966 Sequential estimation of quantal response curves: A new method of estimation. *Biometrika.* 53: 439-454.
- Yam, J., Reer, P.J. & Bruce, R.D. 1991 Comparison of the up-and-down method and the fixed dose procedure for acute oral toxicity testing. *Fd. Chem. Toxicol.* 29:259-263.

Computer Simulations in Study design

Statistical simulations allow us to determine the accuracy of the test design in estimating LD50 in ways that would not be possible with a single sample or even a small number of samples run in actual animals. Since the laboratory to laboratory and intra laboratory variability is not different with the new test designs, the only question is how well they can accurately predict the 'true' values.

Prediction of the 'true' LD50 for a population of rats will depend both on the size of the sample of the population that is sampled, the degree of variability of the response with the population of rats, and the statistical method that is used to estimate the result. Because the LD50 test results in a simple yes/no answer, it is possible to use computers to simulate the degree to which any specific statistical procedure can estimate the 'true' LD50 of the population.

Simulations are done in a stepwise fashion. First the 'true' result is assigned to a 'virtual population' of rats, secondly the population is assigned a known or 'true' degree of variability (or slope of the dose response curve). Because the simulations are being run on a computer, a very large number of 'virtual populations' can be defined each with a different combination of 'true' LD50 and 'true' slope. Simulations can be done for any, (and as many as desired) combinations of 'true' LD50 and 'true' slope as the investigator is willing to simulate. This allows for very rigorous examination of the robustness of the statistical procedures that would not be possible in animal studies.

Once the 'virtual population' is defined, the computer picks animals at random from the population as the sample that would be chosen for the actual test. For each animal the computer, based on the probabilities assigned to the 'virtual population', assigns where it will die on the dose response curve. These probabilities are based on normal statistical estimates of population responses. This mimics exactly what happens in actual practice where the study director picks a small number of animals at random to run his or her test each of which has a built in biological variability. The only difference is that the study director only runs the test with one sample or possibly two samples from the populations and assumes that samples were representative of the full population. The computer on the other hand, can pick random samples over and over again and determine how often the test design used will accurately estimate the 'true' LD50 of the population. For instance, in the simulations that were done for the UDP, between 2500 and 10,000 different random samples were picked from each well-defined population of rats. The results of these simulations provide statistical values on the chance that any one random sample of animals will accurately be able to predict the 'true' LD50 of the population. This information is not available if only one random sample is examined via an actual animal study.

One question has been whether a computer simulations isn't 'too' perfect in that the simulated animals will always give results that fit within the assigned parameters for their 'virtual population'. Using simulations it is possible to address this issue by setting up the computer runs to include one, or more animals, that do not respond correctly. For instance, EPA has calculated the ability of one of the 8 test designs to accurately predict the LD50 if the first animal dies independently of whether this was the 'correct' response for that animal. These questions could

not easily be answered by actual animal studies since it would be impossible for the study director to know that the result from the first animal was not predictive of the 'true' population.