

Alternative Methods for the Median Lethal Dose (LD₅₀) Test: The Up-and-Down Procedure for Acute Oral Toxicity

*Amy Rispin, David Farrar, Elizabeth Margosches, Kailash Gupta, Katherine Stitzel, Gregory Carr,
Michael Greene, William Meyer, and Deborah McCall*

Abstract

The authors have developed an improved version of the up-and-down procedure (UDP) as one of the replacements for the traditional acute oral toxicity test formerly used by the Organisation for Economic Co-operation and Development member nations to characterize industrial chemicals, pesticides, and their mixtures. This method improves the performance of acute testing for applications that use the median lethal dose (classic LD₅₀) test while achieving significant reductions in animal use. It uses sequential dosing, together with sophisticated computer-assisted computational methods during the execution and calculation phases of the test. Staircase design, a form of sequential test design, can be applied to acute toxicity testing with its binary experimental endpoints (yes/no outcomes). The improved UDP provides a point estimate of the LD₅₀ and approximate confidence intervals in addition to observed toxic signs for the substance tested. It does not provide information about the dose-response curve. Computer simulation was used to test performance of the UDP without the need for additional laboratory validation.

Key Words: acute oral toxicity; alternative test methods; computer simulation; reduction; sequential testing; staircase design; up-and-down procedure; validation

Introduction

Sequential test designs, in contrast to test designs with fixed sample size involving replicate sampling, can achieve efficiencies in the number of test samples needed by sampling one or a few test subjects at a time until just enough measurements are made to evaluate the experi-

mental endpoint of concern with the desired precision. In general, sequential designs can be applied to either qualitative (yes/no) or quantitative outcomes; however, only the former are considered in this article. Staircase designs, which are a form of sequential test design, permit trials to converge rapidly on the region of interest, such as the median effective dose (ED₅₀¹) or the median lethal dose (LD₅₀¹) in a toxicity test. For suitable applications in toxicology, the use of sequential design and nontraditional calculation methods can lead to reduction of animal usage while maintaining the ability of the test to measure desired experimental results.

Acute toxicity testing, which measures the adverse effects that occur within a short time of administration of single dose of a chemical, is one such candidate for the application of sequential design. Such studies, performed principally in rodents, provide information on the health hazards likely to arise from short-term exposure and are usually an initial step in the evaluation of the toxic characteristics of a substance for both health and environmental effects. Acute testing can be used to identify doses associated with target organ toxicity and lethality that may be referable to humans. It serves as the basis for hazard classification and labeling of chemicals and can provide information for comparison of toxicity and dose-response among chemicals. Acute toxicity data may also provide information about the mode of toxic action of a substance, which can aid in the diagnosis and treatment of toxic reactions. It is used to standardize biological products and can serve to establish dosing levels for repeated dose studies. Acute oral toxicity in the rat is also used to determine the level of lethality to terrestrial mammals.

Background

In the past, acute toxicity test methods were designed to provide robust characterization of the dose-response curve by using several animals (usually 5 of each sex) at each of three to five test doses in an effort to measure tolerance to potentially lethal doses of chemicals in a test population of

Authors with the US Environmental Protection Agency (EPA), Washington, D.C., are Amy Rispin, Ph.D., Senior Scientist; David Farrar, M.S., Statistician; William Meyer, B.S., Scientist/Technical Editor; Deborah McCall, B.S., Chief, Technical Review Branch, Registration Division; and Elizabeth Margosches, M.P.H., Ph.D., Statistician, Risk Assessment Division, Office of Pollution Prevention and Toxics. With the US Consumer Product Safety Commission, Washington, D.C., are Kailash Gupta, D.V.M., Ph.D., Veterinary Medical Officer; and Michael Greene, Ph.D., Mathematical Statistician. With The Procter & Gamble Company, Cincinnati, Ohio, are Katherine Stitzel, D.V.M., Associate Director of Human Safety; and Gregory Carr, Ph.D., Principal Scientist, Biometrics and Statistical Sciences Department. This manuscript does not reflect the official policy of the governmental agencies or the company listed.

¹Abbreviations used in this article: ASTM, American Society for Testing and Materials; ED₅₀, median effective dose; EPA, Environmental Protection Agency; ICCVAM, Interagency Coordinating Committee on the Validation of Alternative Methods; LD₅₀, median lethal dose; OECD, Organisation for Economic Co-operation and Development.

laboratory animals. Each group of animals was given a single dose of a chemical, with doses chosen to bracket the expected LD₅₀ (the dose at which 50% of the animals are expected to die). Animals were observed for 14 days with observation of the onset, nature, severity and reversibility of toxicity, as well as the timing of lethality after acute chemical exposure. Ideally, the test would have included doses close to the LD₁₃ and the LD₈₇ doses to obtain the best information. At the least, each dose would have shown a decreasing proportion of survivals with increasing dose levels, and two or more doses would have shown partial responses (Litchfield and Wilcoxon 1949).

To discuss the testing procedures, it is important to define the related terms. (These toxicological terms are largely standardized through long usage described [e.g., Chan and Hayes 1994]). When exposed to a test chemical, the sensitivity of the animals to the lethal effect of the chemical is generally lognormally distributed. The distribution is characterized by two parameters, mu (μ) and sigma (σ), which locate the center of the normal curve and its standard deviation, respectively (Figure 1). Sampling this population in an acute toxicity test allows the *lethality dose-response (or cumulative response) curve* to be characterized (Figure 2). A standard practice in toxicology has been to convert the cumulative response curve to a straight line by a *probit transformation*, where the *slope* is the inverse of σ of the lognormal curve, and the *intercept* is the point estimate of the LD₅₀ and corresponds to μ (Finney 1971; for additional details, Chan and Hayes 1994; Riggs 1976). *Confidence intervals* can be calculated as a function of the sample variance and the number of animals tested to provide plausible bounds on the LD₅₀ location and slope.

The traditional acute toxicity test was designed to yield data for all of the regulatory applications described above. The LD₅₀ and slope could be determined simply by graphing the results if the dose range included at least two doses

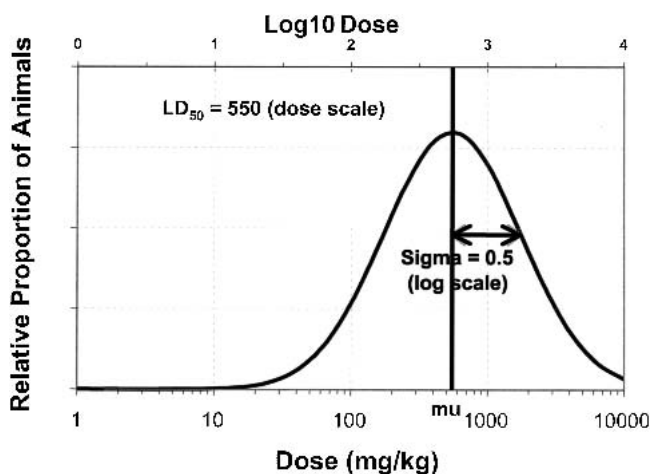


Figure 1 Parameters for tolerance distribution. This particular value (sigma = 0.5) has been chosen for the purpose of illustration.

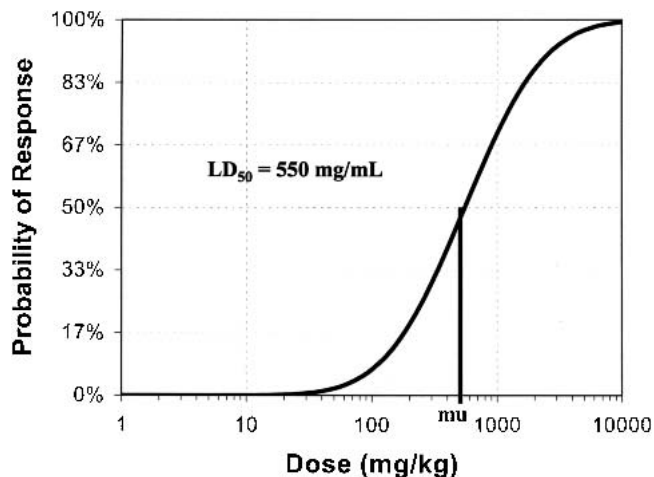


Figure 2 Dose-response curve.

with partial responses. Testing groups of animals at each dose level allowed outliers to be more readily detected. Robust information on the dose-response characteristics of the chemical in the test population allowed authorities to examine the low toxicity end of the curve including estimation of LD_x and confidence interval values for the different assessments performed for regulatory purposes.

However, as originally designed, including range finding for best placement of doses in the main test, the traditional LD₅₀ test could use 50 or more animals. The Organisation for Economic Co-operation and Development (OECD¹) modifications of the test method for use with industrial chemicals, pesticides, and their products called for fewer animals, normally 20 to 30 (OECD 1981, 1987); however, performance of these modified test protocols was not fully examined. In light of the heavy use of animals in traditional test methods, Bruce (1985) applied the up-and-down procedure (UDP¹), a sequential test design, to acute toxicity testing. The use of sequential testing design in an UDP or staircase design was first explored for munitions testing by Dixon and Mood (1948). Scientists in Europe have also incorporated elements of sequential testing in two other alternatives for the traditional LD₅₀ test, namely the fixed dose procedure and the acute toxic class method (OECD 1992, 1996). These two alternative methods provide range estimates of lethality and are applied primarily to hazard classification.

The classical acute toxicity test design is sometimes inefficient. Dixon (1991) noted,

Classical designs center test levels around [the ED₅₀] value . . . A poor guess will result in some tests being conducted at levels that contribute little information to the ED₅₀ estimate . . . If all animals are tested at once using a classical test design based on a poor guess, a large fraction of them may be tested at levels remote from the ED₅₀ . . . [U]ncertainty about the location of the ED₅₀ or about the variance of the threshold distribu-

tion [the distribution of animal responses], makes selection of an optimal number of levels and an optimal number of animals per level for a classical design difficult. Unlike classical designs, sequential designs can quickly correct for a poor guess . . . [p 16-17]. Sequential designs...[permit] trials to converge on the region of the level to be estimated, and thus provide more efficient estimates of this level [p 6].

In most statistical experiments, the sample size is fixed in advance based on some knowledge about the process and the desired statistical properties of the estimate or hypothesis test that was the reason for the experiment. An experiment based on a sequential design is different in that it is conducted one or a few observations at a time. Sequential designs protect the experimenter when little is known about the process because the actual sample size and allocation to type and amount of treatment can change during the experiment. The experimenter monitors the results as they occur to determine whether there are enough data to perform the required analysis or whether additional observations are needed. When little is known, fixed sample sizes may not have the right properties, consume too many resources, or both.

The UDP for acute oral toxicity involves testing young adult animals one at a time in a staircase fashion. Unless the conditions for terminating the test have been reached, the next dose level depends on the results in the previous animal. If that animal survived, then the next dose is higher, whereas if the last animal died, then the next dose is lower. This process permits the experiments to converge on the region of the LD₅₀, which is the inflection point corresponding to the median response of the lognormal dose-response curve. The technique was evaluated in laboratory studies in 1987 (ICCVAM 2001). Subsequently, two other studies were done comparing the results of the UDP with those using traditional LD₅₀ test methods, for 35 substances. As reported in Lipnick et al. (1995), the results revealed that the UDP was able to predict the LD₅₀ of these materials as well as the traditional test. The up-and-down study design for acute toxicity was accepted as a standard test method by the American Society for Testing and Materials (ASTM¹) in 1987 (and later updated in 1998) and by OECD in 1998 (OECD guideline 425) as an alternative for the acute oral toxicity guideline OECD guideline 401 (1987) using traditional fixed sample design.

However, the ASTM (1987) and OECD 425 (1998) acute oral toxicity UDP guidelines had been designed for use with certain types of chemicals that tended to have steep dose-response slopes (i.e., ≥ 8). These guidelines used an assumed value of σ (corresponding to a steep slope of the cumulative LD₅₀ curve) for setting the dose progression and calculating confidence intervals. The test performed best with initial knowledge of approximate LD₅₀ and slope. However, the use of a sighting study could call for additional animals. Computer simulations (ICCVAM 2001) have shown that poor choice of starting dose could lead to the use of many animals before the test converges, or could

introduce a significant bias toward the starting dose in the LD₅₀ estimate. The use of assumptions about σ and slope introduced additional inaccuracies when estimating confidence intervals unless the slope and σ of the actual population corresponded closely to the assumed values. Finally, given the built-in assumptions of small σ and steep slope, the test performed best for chemicals with steep slopes.

In 1999, the OECD called for the three alternative guidelines: the fixed dose procedure (OECD 420; OECD 1992), the acute toxic class method (OECD 423; OECD 1996), and the UDP (OECD 425; OECD 1998) to be revised so that they could be used to replace the traditional acute toxicity test. They agreed that "These changes will . . . [take] advantage of sequential dosing and appropriate statistical methods, and include the incorporation of and use of data from well-designed sighting studies. These changes should, to the extent feasible, reduce the number of animals used and introduce refinements to reduce the pain and distress of animals" (OECD 1999, p. 1). The United States, which had sponsored the use of the UDP as an alternative to the traditional acute toxicity test, redesigned this test to meet the OECD mandate. The design team consisted of scientists and statisticians from the Environmental Protection Agency (EPA¹), the Consumer Product Safety Commission, and The Procter & Gamble Company.

Revised UDP

Development of the Revised UDP

The goals of the US design team were to fulfill the OECD mandate and to develop a robust test that would be applicable to a wide variety of pesticides, industrial chemicals, and their products. The test should have the ability to be performed, and for results to be calculated, without the need to assume a value for the standard deviation, σ . Computer simulations were used to evaluate the performance of the previous version of the OECD guideline (OECD 425, 1998) and to determine appropriate changes to optimize the method's performance without actually testing animals in the laboratory.

The revised test guideline (Figure 3) was peer reviewed by independent panels of experts convened by the Inter-agency Coordinating Committee on the Validation of Alternative Methods (ICCVAM¹) (ICCVAM 2001) and by the EPA Scientific Advisory Panel (EPA 2001). The main test and the limit test are described below.

Performance of the Revised UDP

The revised OECD guideline 425 (OECD 2001a) has improved performance for prediction of the point estimate of lethality (LD₅₀) and confidence intervals for chemicals with wide variability of response characteristics, even when the approximate LD₅₀ and dose-response slope are not known.

MAIN TEST

- Default dose progression factor
3.2 × dose
- Flexible stopping rule
to allow for variations in *sigma* or slope
- Confidence interval calculation
- Default starting dose of 175 mg/kg
- AOT 425 Pgm software

LIMIT TEST

- Sequential dosing
- 2000 mg/kg—5 animals
- 5000 mg/kg—3 to 5 animals

Figure 3 Up-and-down procedure.

In addition, the revisions allow it to be used to evaluate lethality in the 2000 to 5000 mg/kg range for certain hazard classification purposes. The efficiencies of the new up-and-down dosing routine allow the LD₅₀ to be estimated using relatively few animals. However, the guideline does not provide for determination of the slope of the dose-response curve. (When available from other acute tests, the slope is used to calculate other LD_x values for use in human and environmental risk assessments and to improve assessment of the contribution of the substance to the toxicity of a mixture so that it is not necessary to test the mixture itself in animals.)

UDP Guideline

The guideline is available from OECD (as guideline 425) and through the EPA as Office of Prevention, Pesticides, and Toxic Substances Harmonized Test Guideline 870.1100 Acute Oral Toxicity (Draft). The EPA guideline and full documentation of its performance is available on an EPA website (EPA 2002). Characteristics of the new UDP guideline are summarized here and described in more detail below. It uses sequential dosing and calls for testing to be performed in a single sex to reduce variability in the test

population. Animals are tested individually in a staircase fashion using a dose progression and starting dose based on characteristics of the chemical being evaluated. A flexible or adaptive stopping rule limits the number of animals in the main test while allowing the method to be applied to chemicals with a wide range of slopes of the dose-response curve. A computer program is available to determine whether conditions for terminating the test have been reached. Guidance is provided for use of all available information to determine initial dosing and dose progression spacing. Initial doses should be set at sublethal levels, and dose progression should be based on an approximate slope. However, the guideline includes recommendations for default values for starting dose and dose progression for use in the absence of initial information. As in the traditional acute test, this replacement UDP provides for clinical observations over 14 days.

A maximum likelihood method is used to estimate median lethal dose or LD₅₀. Profile likelihood methods are used to estimate confidence intervals. The LD₅₀ estimate is obtained using the assumed value of σ , which sets the dose progression (in default, a slightly more moderate one than in the past). The confidence intervals, however, have broader applicability than before and do not require assumptions about σ . For substances for which insufficient information is available to provide a prior estimate of slope and LD₅₀, the main test incorporates elements of range finding by using widely spaced doses. Although the statistical procedures are more complicated than those for a nonsequential test, the use of computer software can simplify much of the complexity.

In addition, a sequential limit test that uses up to five animals has been substituted for the 10-animal batch limit test. Sequential observations permit a separate decision about the contribution of each animal to the determination of whether the LD₅₀ is above or below the limit dose. The same principle contributes to the fixed dose procedure and acute toxic class method limit tests.

Use of Statistical Simulation for Development and Validation of the UDP

By varying assumed values of the LD₅₀ and σ for a hypothetical set of animals exposed to a toxic chemical, it is possible to have a computer generate responses from a randomly chosen sample of the hypothetical population. A computer can use these hypothetical responses to simulate the results from thousands of small samples of the underlying population. Because the underlying mean and standard deviation of the test population are known, these simulations then can be used to determine whether changes in the test design would improve the ability of the UDP test to estimate the mean and standard deviation. By varying the standard deviation assigned to the hypothetical set of animals, it is possible to simulate the degree of variation in the population's response that would occur because of animal-

to-animal, inter- and intralaboratory species, strain, sex, age, and housing variability. Such simulations have shown that the new sampling technique used in the UDP has a much better chance of placing the estimated LD_{50} close to the value that was used in defining the underlying hypothetical population even when the starting dose is inappropriate (ICCVAM 2001, Appendix 0-2). This type of comparison would not be practical using actual animal tests because it would be impossible to determine whether each small sample tested is providing correct or incorrect estimates of the underlying population.

Actual animal testing was not necessary for determination of the validity of the new statistical design. It is not appropriate or possible to compare changes in sampling technique or to assess the ability of a new statistical design to accurately estimate the mean and standard deviation of the population based on the results of a few runs of a test because there is no way to determine the goodness of fit of the statistical procedure from a few samples. However, computer simulations can be used to compare the results of thousands of individual hypothetical tests. By using a large series of such simulations and varying starting dose, dose progression, and assumptions about μ and σ of the underlying population, it is possible to test how often a new statistical sampling technique will accurately estimate the LD_{50} and σ , or standard deviation, of the population.

Stopping Rule

Simulations revealed that the number of test subjects needed to provide an acceptable degree of accuracy depends, in part, on the slope of the dose-response curve of the test population. However, in most cases, the slope is not known in advance of testing. Therefore, to allow the up-and-down method to be applied to a wide variety of chemicals with reasonable reliability, a flexible stopping rule using criteria based on an index related to the statistical error was developed and incorporated into the test. The testing stops when one of the following stopping criteria is met (EPA 2001; ICCVAM 2001):

1. Three consecutive animals survive at the upper bound of dosing, where the upper bound of testing is the highest dose given to the animals, based on regulatory needs. Normally, the upper bound of dosing is 2000 or 5000 mg/kg.
2. Five reversals occur in any six consecutive animals tested;
3. At least four animals have followed the first reversal and two likelihood ratios, which compare the LD_{50} estimate with LD_{50} values above and below (Figure 4), exceed a critical value of 2.5. In this example, a likelihood is a probability of seeing a certain pattern of survivals or deaths based on hypothetical combinations of LD_{50} and slope consistent with the data.

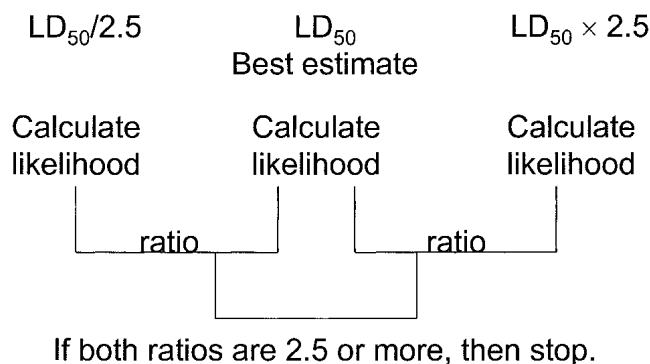


Figure 4 Likelihood ratio stopping rule.

For chemicals with higher slopes, the stopping rules will be satisfied with four animals after the first reversal. Additional animals may be needed for chemicals with dose-response slopes below 4. In the interest of animal welfare, testing in any case is terminated at 15 animals. Simulations have shown that the average animal use is expected to be seven to nine animals per test (Westat 2001).

Choice of Starting Dose and Dose Progression Factor

As noted above, the revised UDP works best when the approximate LD_{50} and slope of the test chemical are known. Before conducting the study, the testing laboratory should consider all available information on the test substance including in vitro data or test results for similar substances to select the best initial dose and dose progression or spacing (Rispin et al. 2002). Ideally, the initial dose should be just below the prior estimate of the LD_{50} of the test material and dose progression should be based on the toxicologist's best estimate of slope of the dose-response curve, where slope = $1/\sigma$ (Figure 2 and definitions in the Introduction). Most industrial chemicals tend to have slopes of 8 or higher. Many pesticides and other chemicals whose toxicity is receptor mediated may exhibit slopes as low as 2.0 or 2.5.

In the absence of initial information indicating likely slope or LD_{50} , the UDP guideline recommends a default starting dose of 175 mg/kg and the use of half log units of dose corresponding to a progression of 3.2. These default values will accommodate a variety of situations, including chemicals with slopes as low as 2.0. In addition, 175 mg/kg has been chosen because it is likely to be below the LD_{50} of most chemicals.

Although flexible stopping rules allow the UDP to be applied to test materials with a wide range of slopes, for optimum performance of the UDP, the dose progression used should be based on an accurate estimate of σ . In addition, to account conservatively for any bias in the LD_{50} estimate, it is essential to initiate dosing at a dose below the actual LD_{50} value. Setting initial doses at sublethal levels

also ensures that LD₅₀ values are not underestimated while reducing distress in the animals (Westat 2001). The following two cases describe the outcome when an accurate estimate of σ is not available.

Example 1: Assumed $\sigma \ll$ true σ

When the assumed σ (i.e., the σ on which the dose progression is based) is much smaller than the true σ of the actual test population, the estimated LD₅₀ may be biased in the direction of the starting dose. For example, if the starting dose is less than the true LD₅₀ of the test population, the estimated LD₅₀ will generally be below the true LD₅₀. Also, if the starting dose is greater than the true LD₅₀ of the test population, the estimated LD₅₀ will tend to be greater than the true LD₅₀. To minimize the chance of overestimating the LD₅₀ due to this bias, the UDP guideline recommends a choice of starting dose just below the assumed LD₅₀.

Example 2: Assumed $\sigma \gg$ true σ

If the assumed σ on which the dose progression is based is much larger than the true σ of the test population, the median estimated LD₅₀ can be much larger or much smaller than the true LD₅₀ depending on the starting dose. In this case, the LD₅₀ can be estimated only within a range.

A software program has been developed by EPA to assist the user in setting test doses, to determine when the stopping rules have been satisfied, and to calculate the LD₅₀ and confidence interval (Westat 2001). The software will run on a personal computer, and it simplifies the use of this test method for the toxicologist.

Calculation of the LD₅₀ Estimate

The LD₅₀ estimate is calculated using the maximum likelihood method unless the response pattern falls into an exceptional case. All deaths, whether immediate or delayed, or due to humane sacrifice, are incorporated for the purpose of the maximum likelihood analysis.

In performing the maximum likelihood calculation, an estimate of σ of 0.5 (corresponding to a slope of 2) is used unless a better generic or case-specific value is available. This default value of dose spacing of antilog (0.5), where $(\log(3.2) = 0.5)$, is larger than in previous versions of the test; it was found to permit the test to perform better over a wider range of substances. If a better value of σ is available, the dose spacing should be adjusted accordingly before the test is initiated. Because the accuracy of the estimated LD₅₀ improves as the estimated σ (used to set the dose progression) approaches the true σ of the underlying population, the toxicologist should make every effort to improve the accuracy of the estimated σ .

Under some circumstances, statistical computation of

the LD₅₀ will not be possible or will likely give erroneous results. (Table 1 provides a tabulation of all possible test outcomes.) Special means to determine or report an estimated LD₅₀ are available for these circumstances as follows:

1. If testing is stopped because an upper or lower boundary dose was tested repeatedly, then the LD₅₀ is estimated to be greater than the upper bound or less than the lower bound as appropriate.
2. If all of the dead animals have higher dose levels than all of the live animals, then the LD₅₀ is between the doses for the live and the dead animals, with reasonable confidence. These observations give no additional information on the exact value of the LD₅₀ (Blodgett 1997). Still, a maximum likelihood LD₅₀ estimate can be made, provided there is a value for σ . When the actual value of σ is not available, the computer program provided calculates a maximum likelihood estimate based on the dose spacing used. However, when this type of response pattern is seen, it is clear that the actual σ is much smaller than the value used to set the dose progression. Therefore, point estimate LD₅₀ values are artificial. Rather, the yield of the test is a *range* for lethality. If a closely related substance is tested, testing should proceed with a smaller dose progression.
3. If the live and dead animals have only one dose in common, all of the other dead animals have higher doses, and all of the other live animals lower doses, then the LD₅₀ equals their common dose. If a closely related substance is tested, testing should proceed with a smaller dose progression.

If none of the above situations occurs, then the LD₅₀ estimate is calculated using the maximum likelihood method (ICCVAM 2001, Appendix K, The UDP Primary Test: Proposed Revision of the Guideline 425).

Confidence Interval

The confidence interval, which is based on the entire experimental data set from the main test, provides information on the repeatability of the estimate. Confidence intervals can be used to provide the regulator with a basis for evaluating how to incorporate test results into regulatory applications. The confidence interval is a gauge of the capacity of the data collected to give information on the spread of possible responses, and it provides plausible bounds on the locations of the LD₅₀ values.

A class of methods called *profile likelihood* may be used to calculate confidence intervals consistent with the experimental data (Meeker and Escobar 1995). The likelihood is a certain function of the data (in this case, the pattern of survivals or deaths in response to the chemical), which provides an index of support that the data provide for alternative combinations of parameters (in this case, the LD₅₀

Table 1 Outcomes of the up-and-down procedure: cases and confidence intervals

Case no.	Definition of case	Proposed approach	Possible findings
1	No positive dose-response association. 1a) all animals tested in the study responded, or 1b) none responded, or 1c) the geometric mean dose is lower for animals that responded than for animals that did not respond.	LD ₅₀ ^a cannot be calculated. Confidence interval not applicable.	Possible inferences: 1a) LD ₅₀ < lowest dose; 1b) LD ₅₀ > highest dose; 1c) reverse dose-response curve; unlikely test outcome. In case 1b, the highest dose tested is equivalent to a limit dose.
2	Multiple partial responses. One or more animals responded at a dose below some other dose where one or more did not respond. The conditions defining Case 1 do not hold. (The definition of Case 2 holds if there are 2 doses with partial responses, but holds in some other cases as well.)	Maximum likelihood estimate and profile likelihood computations of confidence interval are straightforward.	The LD ₅₀ can be estimated and its confidence interval calculated.
3	No intermediate response fractions. One or more test doses is associated with 0% response and one or more is associated with 100% response (all of the latter being greater than all of the former), and no test doses are associated with a partial response.	Lower bound = highest test dose with 0% response. Upper bound = lowest test dose with 100% response.	High confidence that the true LD ₅₀ is between the two bounding doses. Any value of LD ₅₀ between highest dose with 0% response and lowest dose with 100% response is equally plausible.
4	One partial response fraction, first subcase. An intermediate partial response is observed at a single test dose. That dose is greater than doses associated with 0% response and lower than doses associated with 100% response.	The LD ₅₀ is set at the single dose showing partial response and its confidence interval is calculated using profile likelihood method.	The LD ₅₀ can be estimated and its confidence interval calculated.
5	One partial response fraction, second subcase. There is a single dose associated with partial response, which is either the highest test dose (with no responses at all other test doses) or the lowest test dose (with 100% response at all other test doses).	The LD ₅₀ is set at the dose with the partial response. A profile likelihood confidence interval is calculated and may be finite or infinite.	The true LD ₅₀ could be at the boundary of the testing range with more or less confidence.

^aLD₅₀, median lethal dose.

value and slope). Such a calculation of profile likelihood confidence intervals requires calculating the profile likelihood for different values of fixed assumed LD₅₀ values with their corresponding profile maximizing slopes and finding the value for which the profile likelihood equals a critical value. This procedure, which is computationally intensive, has been incorporated into the software for the guideline.

In traditional LD₅₀ tests, the stated confidence interval (e.g., 95%) is exactly what is calculated. However, for the UDP, the algorithm used to compute 90, 95, or 99% profile

likelihood confidence intervals is not exact but only approximate, so that in some situations, the stated confidence interval will not provide the desired coverage or may provide more than the desired coverage. Here, the “coverage” of the confidence interval is the probability that a calculated confidence interval, based on an acute toxicity experiment, actually encloses the true LD₅₀ for a population. The random stopping rules in the UDP improve the ability of the test to respond to a variety of types of chemical, but this characteristic also causes the reported level of confidence

and the actual level of confidence to differ somewhat (Shiryayev and Spokoiny 2000). Simulations indicate that actual coverage for the nominal 95% confidence interval is less than 95% when the slope is shallow and more than 95% when slopes are very steep. The nominal 95% confidence interval will have coverage of at least 90% if the slope is 2 to 4 or more (σ 0.25-0.5 or less). For most situations, the coverage will be better than 90% if the slope is 2 or greater. Coverage will be 80% or better if the slope is at least 1. For slopes as low as 0.5, the lowest slope assumed in simulations, the coverage may be as low as 70% (Westat 2001). Probably no type of confidence interval would be narrower than the dose spacing.

Depending on the outcome of the test (Table 1), one of the following three different types of confidence interval estimates of the true LD_{50} is calculated:

1. When the UDP provides a point estimate of the LD_{50} . When at least three different doses have been tested and the middle dose has at least one animal that survived and one animal that died, a profile-likelihood-based computational procedure can be used to obtain a confidence interval.
2. When the UDP provides a range estimate of the LD_{50} . If all animals survive at or below a given dose level and all animals die when dosed at the next higher dose level, a confidence interval is calculated that has as its lower limit the highest dose tested where all of the animals survive and has as its upper limit the dose level where all of the animals died. An approximate confidence interval can be obtained; however, because this type of response would ordinarily occur when the dose-response is steep, in most cases the true LD_{50} is expected to be contained within the calculated interval or will be very close to it.
3. When the dose-response curve is flat or the standard deviation is large. In some instances, confidence intervals are reported as infinite and may occur (e.g., when the slope of the dose-response is relatively flat or relatively uncertain).

Performance of the UDP Guideline

Data gathered under the UDP fit into one of five animal response patterns (Table 1 cases 1-5). These results can occur whether the study was carried out using fixed sampling (as in the classical test) or sequential procedures. There are two ways, however, in which the sequential design of the UDP makes consideration of these cases more important: (1) The classical test relied only on the doses with partial responses. With the UDP, as fewer animals are used, it becomes more likely that some doses will be represented by a single animal and consequently show an all-or-none response. (2) To permit fewer animals to be used, it is even more important than for the classical test for as many of the doses as possible to be used in estimation. By

considering outcome configurations, the approaches to calculation of LD_{50} and confidence intervals used in the UDP can make efficient use of the tested animals.

For Cases 1, 2, 4, and 5, a point estimate is obtained. Case 3 has no partial responses; all animals die at higher doses and all animals survive at lower doses. This result implies that the LD_{50} is between the highest dose with no response and the lowest dose where complete responses occur. This case occurs most often when the dose spacing is large relative to the actual standard deviation of the lethality normal curve. In this case, any value between these two doses might be the true LD_{50} and the test response is a range estimate of lethality. In effect, the two doses can serve as a 95% confidence interval.

Prediction of Hazard Classification by the Revised UDP

The international community has harmonized criteria for classification and labeling of chemicals (OECD 2001b). The agreed-upon acute oral LD_{50} cutpoints are 5, 50, 300, 2000, and 5000 mg/kg. Pursuant to the United Nations formal adoption of the Globally Harmonized System in December 2002, competent authorities in all countries are expected to adopt the system.

Simulations verified that the default procedure indicated in the guideline—with an initial test dose of 175 mg/kg, a minimum test dose of 1 mg/kg, a maximum test dose of 5000 mg/kg, and use of a likelihood-ratio stopping rule—provides good performance for the UDP for classification when dose progression spacing of 3.2x dose is used. The dose progression in mg/kg is 1.75, 5.5, 17.5, 55, 175, 550, 1750, 5000. A probit model is assumed. Extensive simulations were performed to evaluate the point estimate and confidence interval outputs of the UDP for classification and for hazard and first tier risk assessment. The results of these simulations can be found on the ICCVAM and EPA web sites (ICCVAM 2001, Appendix L, Comparison of 5 Stopping Rules and 2 LD_{50} Estimators Using Monte Carlo Simulations; Westat 2001).

Simulations have shown that when the default dose progression and a starting dose of 175 mg/kg are used, when a chemical is misclassified, it will be more often assigned to a more toxic category than to a less toxic category. This has to do with the relation between the initial test dose and the category boundaries. The precision of the UDP is limited by the dose progression factor. In particular, in steep-slope situations, the maximum-likelihood estimate may be between two test doses that differ by a factor of 3.2 and may straddle a category boundary. Therefore, chemicals with LD_{50} values within certain intervals may be consistently overclassified or consistently underclassified. For this and other reasons noted above, it is important for the toxicologist to use all available information when determining the dose progression as it relates to actual slope of the chemical.

Limit Test

The purpose of a limit test is much less ambitious than an UDP. The limit test is required only to estimate whether the LD₅₀ is greater or less than a certain value—the limit. Animals are dosed at the limit, the number of deaths is observed, and a decision is made as to whether the LD₅₀ is greater or less than the LD₅₀. The US design team modified the limit test to be more like the main test in that it is executed sequentially. The new limit test is designed to classify at either 2000 mg/kg or 5000 mg/kg (ICCVAM 2001, Appendix M. The UDP Limit Test: Accuracy of In Vivo Limit Dose Tests).

Until recently, the standard procedure for limit dose tests was a 10-animal fixed sample test. The limit test for the UDP uses a sequential sampling design to improve the reliability of correct classification over that obtained from batch testing for a given number of animals. Sequential testing plans classify adequately in comparison with fixed sample plans using up to twice as many animals, particularly when the true LD₅₀ is either much less or more than the limit dose. The classification deteriorates when the true LD₅₀ approaches the limit dose. Classifications are also less accurate when the standard deviation, or σ , of the test population increases.

In Table 2, the probability of correct classification using a five-animal sequential test plan for a limit dose of 5000 mg/kg is shown. Every entry in the table represents the probability that the correct classification would occur given assumed values of the LD₅₀ and σ . The plan is very accurate for low values of LD₅₀ and σ , and it gradually decreases in accuracy with increasing σ and as the LD₅₀ approaches the limit dose. For example, with values of σ at 0.12 or 0.25, the

probability of correct classification is between 99 and 100% for a true LD₅₀ up to 2000 mg/kg. This plan is also biased toward classifying the LD₅₀ below the limit dose, which provides a margin of safety.

The five-animal sequential limit test proceeds as follows:

1. Dose one animal. If the animal dies, the test is concluded and the decision is made that the LD₅₀ is below the limit dose.
2. If the first animal survives, continue testing until either (a) two more animals survive (denoting that the LD₅₀ is above the limit dose), or (b) three animals die (denoting that the LD₅₀ is below the limit dose).

Following this procedure, the limit dose test can conclude with as few as one animal, if the first animal dies, or use as many as five animals, if survivals and deaths alternate. On average, fewer than five animals will be used.

The five-animal sequential test plan compares favorably with a 10-animal fixed test plan. This test is also biased toward classifying the LD₅₀ below the limit dose. At low values of σ and for low LD₅₀ values, there are no practical differences in the probability of correct classifications. Generally the differences are less than 6% under 3000 mg/kg. The probability of correct classification for the 10-animal plan is shown in Table 3.

As would be expected from a plan with fewer animals, the correct classification probabilities for the five-animal sequential test decrease somewhat from the 10-animal plan in Table 3. For low LD₅₀ values, the results are very close between both plans. For values of the LD₅₀s that exceed the limit dose, the sequential plans tend to classify correctly less

Table 2 Probability of correct classification for five-animal sequential test plan (limit dose = 5000 mg/kg)

LD ₅₀ ^a	Sigma				
	0.12	0.25	0.5	1.25	2
1.5	1.00	1.00	1.00	1.00	1.00
50	1.00	1.00	1.00	1.00	0.98
250	1.00	1.00	1.00	0.98	0.93
1500	1.00	1.00	0.98	0.86	0.79
2000	1.00	1.00	0.96	0.82	0.76
3000	1.00	0.97	0.87	0.75	0.72
5000	0.66	0.66	0.66	0.66	0.66
6000	0.71	0.53	0.44	0.38	0.37

^aLD₅₀, median lethal dose.

Note: The decision rule for the five-animal sequential plan is that the LD₅₀ is less than the limit dose if the first animal dies, or if three animals die. The LD₅₀ exceeds the limit dose if three animals, including the first, survive.

Table 3 Probability of correct classification for 10-animal fixed plan (limit dose = 5000 mg/kg)

LD ₅₀ ^a	Sigma				
	0.12	0.25	0.5	1.25	2
1.5	1.00	1.00	1.00	1.00	1.00
50	1.00	1.00	1.00	1.00	1.00
250	1.00	1.00	1.00	1.00	0.98
1500	1.00	1.00	1.00	0.92	0.84
2000	1.00	1.00	0.99	0.87	0.80
3000	1.00	1.00	0.93	0.78	0.73
5000	0.62	0.62	0.62	0.62	0.62
6000	0.92	0.69	0.54	0.44	0.42

^aLD₅₀, median lethal dose.

Note: The decision rule is that the 10-animal fixed test plan classifies the LD₅₀ less than the limit dose if five or more animals die, or greater than the limit dose if fewer die.

frequently than the 10-animal fixed dose plan. This result means that more chemicals would be erroneously considered to have the LD₅₀ below the limit dose. For regulatory purposes, this type of misclassification is probably better than erroneously classifying the LD₅₀ above the limit dose.

For comparison, the five-animal fixed sample test plan is shown in Table 4. The rules of this sequential plan are different from the fixed plans. The sequential plan stops the test with the death of the first animal, which cannot be done with the fixed plans. The result is that the sequential plans are more accurate than the fixed plans when the test uses chemicals that have LD₅₀ values below the limit dose. The fixed plans are more accurate with chemicals that have an LD₅₀ above the limit dose. When the LD₅₀ is very low or very high and sigma is low, both types of tests perform accurately.

Humane Practices

The new OECD 425 guideline provides a significant improvement in the number of animals used compared with guideline 401, which required at least 20 animals in a test. In addition, it contains a requirement to follow the OECD Guidance Document on Humane Endpoints (OECD 2000). This document attempts to reduce pain and distress of animals in toxicity tests by permitting early killing of the animals if a set of clinical signs is identified that can reliably predict the outcome. If the progression of toxic signs is invariable, the use of humane endpoints may be possible in sequential studies. In addition, this OECD guidance document defines a set of clinical signs that may be used to justify early killing of the animals and prescribes frequent observation of animals showing signs of pain and distress.

Table 4 Probability of correct classification for five-animal fixed test plan (limit dose = 5000 mg/kg)

LD ₅₀ ^a	Sigma				
	0.12	0.25	0.5	1.25	2
1.5	1.00	1.00	1.00	1.00	1.00
50	1.00	1.00	1.00	1.00	0.97
250	1.00	1.00	1.00	0.97	0.89
1500	1.00	1.00	0.97	0.78	0.69
2000	1.00	1.00	0.93	0.72	0.65
3000	1.00	0.95	0.80	0.63	0.58
5000	0.50	0.50	0.50	0.50	0.50
6000	0.89	0.72	0.62	0.55	0.53

^aLD₅₀, median lethal dose.

Note: The decision rule is to classify the chemical's LD₅₀ as less than the limit dose if three or more animals die and more than the limit dose if three or more animals survive.

All animals killed for humane reasons are considered in the same way as animals that died on test. In addition, as noted above, initiating dosing below the LD₅₀ minimizes the number of animals killed before the first reversal.

Conclusions

The revised UDP guideline provides a point estimate of LD₅₀ and approximate confidence intervals in addition to observed signs of toxicity. The staircase design used in the guideline test builds in efficiencies by concentrating most doses near the region of the LD₅₀ and by using flexible or adaptive stopping rules to accommodate a variety of chemicals. Careful choice of dose progression and initial dose also improves performance. However, when assumptions about standard deviation of the test population diverge significantly from actual values, point estimates of the LD₅₀ may not be possible. In these cases, a range estimate is provided.

The UDP does not allow for characterization of the slope because efficient management of doses toward the region of the LD₅₀ does not provide enough doses in the wings of the dose-response curve. If slope is needed, another method would be used. The revised UDP guideline can be used for chemicals with a wide variety of actual dose-response slopes and can be used for hazard classification and certain other hazard and risk assessment purposes. The five-animal sequential plan used in the limit test produces results that are almost as good as those for the present 10-animal fixed sample plan, while averaging three to five animals per test. That result represents a substantial reduction of animal subjects over the 10-animal fixed sample plan.

The up-and-down method is one of three new sequential methods recently accepted for use by all OECD countries. The fixed dose method involves dosing groups of five animals in a sequential fashion at a choice from five fixed doses. This method incorporates a sighting sequence and uses signs of toxicity—not death—as an endpoint. The acute toxic class method also uses a choice from the same five fixed doses, but instead of dosing five animals at each step, this method tests either three or six animals at each dose. The acute toxic class method uses death as the endpoint. As recently updated and revised, the fixed dose procedure and the acute toxic class method are designed specifically to allow classification of new materials according to the recently developed Globally Harmonized Classification System (OECD 2001b); they do not provide a point estimate of LD₅₀. The UDP used does not take advantage of the information available on the sequence of events; it uses only the final results. An analysis that reflected the path taken to reach the results would also include the dependencies of dose choices and would be much more complicated. In the future, it may be possible to find a way to use all of the information in a way that will improve the accuracy of the test.

References

- ASTM [American Society for Testing and Materials]. 1987. E 1163-87 Standard Test Method for Estimating Acute Oral Toxicity in Rats. Philadelphia: ASTM.
- ASTM [American Society for Testing and Materials]. 1998. E 1163-98 Standard Test Method for Estimating Acute Oral Toxicity in Rats. Philadelphia: ASTM.
- Blodgett JM. 1997. Adding functions to ensure regression converges. *Theory Methods* 26:1203-1214.
- Bruce RD. 1985. An up-and-down procedure for acute toxicity testing. *Fund Appl Toxicol* 5:151-157.
- Chan PK, Hayes AW. 1994. Acute toxicity and eye irritancy. In: Hayes AW, ed. *Principles and Methods of Toxicology*. 3rd Ed, Chapter 16. New York: Raven Press. p 548-579.
- Dixon WJ. 1991. *Design and Analysis of Quantal Dose-Response Experiments (with Emphasis on Staircase Designs)*. Los Angeles: Dixon Statistical Associates.
- Dixon WJ, Mood AM. 1948. A method for obtaining and analyzing Sensitivity Data. *J Am Stat Assoc* 43:109-126.
- EPA [Environmental Protection Agency]. 2001. Federal Insecticide, Fungicide and Rodenticide Act (FIFRA) Scientific Advisory Panel (SAP) Review of Applicability of the Up and Down Procedure Methodology for Acute Oral Toxicity Testing. Final panel report and EPA presentations available at http://epa.gov/scipoly/sap/2001/index_html.
- EPA [Environmental Protection Agency]. 2002. OPPTS Harmonized Test Guideline 870.1100 Acute Oral Toxicity (Draft); also available at <http://epa.gov/oppfead1/harmonization>.
- Finney DJ. 1971. *Probit Analysis*. 3rd ed. Cambridge: Cambridge University Press.
- ICCVAM [Interagency Coordinating Committee on the Validation of Alternative Methods]. 2001. *The Revised Up-and-Down Procedure: A Test Method for Determining the Acute Oral Toxicity of Chemicals*. Final Report (NIH publication no. 02-4501). Research Triangle Park: NIH/NIEHS. Also available at the ICCVAM web sites (http://iccvam.niehs.nih.gov/methods/udpdocs/udpfin/vol_1.pdf and http://iccvam.niehs.nih.gov/methods/udpdocs/udpfin/vol_2.pdf).
- Lipnick RL, Cotruvo JA, Hill RN, Bruce RD, Stitzel KA, Walker AP, Chu I, Goddard M, Segal L, Springer JA, Myers RC. 1995. Comparison of the up-and-down, conventional LD₅₀ and fixed dose acute toxicity procedures. *Food Chem Toxicol* 33:223-231.
- Litchfield J, Wilcoxon F. 1949. A simplified method of evaluating dose-effect experiments. *J Pharmacol Exp Ther* 96:99-113.
- Meeker WQ, Escobar LA. 1995. Teaching about approximate confidence intervals based on maximum likelihood estimations. *Am Stat* 49:48-52.
- OECD [Organisation for Economic Co-operation and Development]. 1981. *OECD Guidelines for the Testing of Chemicals. Guideline 401 Acute Oral Toxicity*. Paris: OECD.
- OECD [Organisation for Economic Co-operation and Development]. 1987. *OECD Guidelines for the Testing of Chemicals. Guideline 401 Acute Oral Toxicity*. Paris: OECD.
- OECD [Organisation for Economic Co-operation and Development]. 1992. *OECD Guidelines for the Testing of Chemicals. Guideline 420 Acute Oral Toxicity—Fixed Dose Method*. Paris: OECD.
- OECD [Organisation for Economic Co-operation and Development]. 1996. *OECD Guidelines for the Testing of Chemicals. Guideline 423 Acute Oral Toxicity—Acute Toxic Class Method*. Paris: OECD.
- OECD [Organisation for Economic Co-operation and Development]. 1998. *OECD Guidelines for the Testing of Chemicals. Guideline 425 Acute Oral Toxicity—Up-and-Down Procedure*. Paris: OECD.
- OECD [Organisation for Economic Co-operation and Development]. 1999. 29th Joint Meeting of the Chemicals Committee and the Working Party on Chemicals. Paris: OECD.
- OECD. [Organisation for Economic Co-operation and Development]. 2000. *Guidance Document on the Recognition, Assessment and Use of Clinical Signs as Humane Endpoints for Experimental Animals Used in Safety Evaluation. Environmental Health and Safety Monograph Series on Testing and Assessment no. 19*. Paris: OECD.
- OECD [Organisation for Economic Co-operation and Development]. 2001a. *OECD Guidelines for the Testing of Chemicals. Guideline 425 Acute Oral Toxicity—Up-and-Down Procedure*. Paris: OECD.
- OECD. [Organisation for Economic Co-operation and Development]. 2001b. *Harmonized Integrated Classification System for Human Health and Environmental Hazards of Chemical Substances and Mixtures. Document ENV/JM/Mono(2001)6*.
- Riggs DS. 1976. *Substrate-enzyme and drug-receptor interactions. The Mathematical Approach to Physiological Problems: A Critical Primer*. Cambridge MA: The MIT Press. p 272-298.
- Rispin AS, McCall D, Farrar D, Margosches E, Gupta K, Stitzel K, Carr G, Greene M, Meyer W. 2002. *Practical Guidance for Implementation of Alternative Methods for Acute Oral Toxicity. Lab Anim (In Press)*.
- Shiryaev AN, Spokoiny VG. 2000. *Statistical inference for autoregressive models of the first order asymptotic theory*. In: *Statistical Experiments and Decisions. Vol 8, Chapter 5*. London: World Scientific Publishers.
- Westat Incorporated. 2001. *Acute Oral Toxicity Software Program: AOT425StatPgm; User Documentation for the AOT425StatPgm Program; Simulation Results for the AOT425StatPgm Program; and QA Testing for the AOT425StatPgm Program*. Report prepared for EPA under Contract 68-W7-0025, Task Order 5-03. Rockville, MD. The program and User's Manual are available on the EPA web site at http://www.epa.gov/scipoly/sap/2001/index_html. The Simulation Results and QA Testing documents are available on the ICCVAM web site at http://iccvam.niehs.nih.gov/methods/Up-and-Down/Proceduredocs/Up-and-Down/Procedurerpt/Up-and-Down/Procedure_ciprop.htm.