Statistical Basis for Estimating Acute Oral Toxicity
Comparison of OECD Guidelines 401, 420, 423 and 425

Introduction

This document serves to provide short summaries of the scientific basis for each of the four acute oral toxicity tests. It will attempt to describe the statistical strengths and limitations of the various methods for accurately determining a point estimate of the LD50, slope of the dose-response curve for LD50, confidence limits around the point estimate of LD50 and the slope, a point estimate of an LD10 and information on the dose-effect response. In this context, a dose-response curve applies to the estimation of lethality and a dose-effect response applies to the estimation of the change in the variety and distribution of all other types of toxicological signs with the change in dose.

By design not all of the guidelines will provide estimates for all of these endpoints. However, in the context of the comparison of the four tests, it was felt that a detailed comparison of the four methods was warranted. This document is still in draft form and will be finalized after the meeting.

Because the response of a test population to a chemical is influenced by the choice of test species and strain, test conditions, and age, sex or body weight of the animals, the LD50 is commonly described as the lethal response of a compound in a particular population under a discrete set of experimental conditions. As a result, the LD50 values, along with slope and confidence intervals are not absolute, but rather provide a relative index of xenobiotic response for comparison of chemicals. Of course, a similar statement would apply to quantitative endpoints of most laboratory animal toxicology tests. For that reason, test guidelines seek to standardize test conditions, to the extent feasible. A well standardized acute test provides a sound method for comparing acute sensitivity to toxic chemicals.

What follows is a brief description of the motivation for and the mathematical and biological principles underlying each acute oral toxicity method followed by a listing of how each test estimates or does not estimate the specific parameters mentioned above. This document is a supplement to the larger guidance document prepared for the OECD meeting and only covers these points. The larger document should be consulted for a complete description of each test and comparisons of the other benefits and weaknesses of each method. Statistical simulations of all four tests will be presented at the meeting.

Acute Oral Toxicity, Guideline 401

A. Principles underlying the test method: Guideline 401 (1987) is an altemative to the 1981 version incorporating provisions for reduction and refinement. The current guideline calls for a test chemical to be administered to the test population in three positive dose levels, generally spaced logarithmically such that they will span the expected 10% to 90% mortality levels. Dose levels may be based on results from a range-finding study. In the main study, groups of 5 animals of a single sex are tested at each dose. After completion of the study, a single group of animals of the opposite sex is tested.

As a traditional acute oral toxicity test, guideline 401 is based on the fact that lethality is a quantal response. Its measurement will give rise to a frequency distribution of responses reflecting the composite tolerances of the test population upon exposure to graded doses of the test chemical. In practice, most chemicals give rise to an approximately lognormal distribution of deaths versus dose, skewed toward hypersensitivity. When this frequency population is transformed to a logarithmic abscissa, a (symmetric) normal distribution generally results that can be characterized by two parameters, the median and the standard deviation, σ. The median is the dose at which 50% of the animals are killed by the test chemical and is called the LD50. Not all animals will react in the same way to the chemical and thus σ represents the square root of the variance of the test population's response to the chemical. The dose-response curve is sigmoidal in nature and represents the cumulative response of the test animals to the chemical. The inflection point of this sigmoidal curve coincides with the LD50 for the test population.

To analyze the data from test guideline 401, the dose response curve can be linearized by transforming the percentage response for log dosage to probits. The slope, ß, of the transformed dose response curve is 1/σ. Responses can be analyzed by probit analysis (1) which calculates the maximum likelihood fit of the probit log dose line by an iterative weighted linear regression method. This can also be done graphically.

B. Point estimate of LD50: Probit analysis of mortality provides a point estimate of the LD50 provided there are at least two doses with mortality rates not equal to 0% or 100%.

C. Confidence limits on the estimate of LD50: The method of probit analysis can provide interpretive statistics such as the 95% confidence interval of the LD50 in this case.

D. Estimate of the slope of the dose-response curve for lethality Guideline 401 provides the slope of the dose-response curve as a study endpoint providing there are at least two doses which have mortality rates not equal to 0% or 100%.

E. Confidence limits on the slope of the dose-response curve for lethality Confidence limits for the slope of the dose-response curve can be calculated if a slope can be determined.

F. Dose-effect curve for the LD50 Toxic signs and pathology results are measured for the animals in each dose level. Thus, a dose-effect curve can be calculated for specific effects observed if they are quantal provided there are at least two doses in which the effect was not present in either 0% or 100% of the animals.. However, not all effects are quantal and some analysis additional to the probit may be needed to estimate the extent and shape of dose-effect curves.

G. Point estimate of LD10: Guideline 401 can provide a point estimate of the LD10 if a slope of the dose-response curve can be determined.

Fixed Dose Procedure, Guideline 420

A. Principles underlying the test method: The Fixed Dose Procedure (FDP) is a method for assessing acute oral toxicity that involves the identification of a dose level that causes evidence

of non-lethal toxicity (termed *evident* toxicity) rather than a dose level that causes lethality. The method was first suggested by the British Toxicology Society in 1984 (2) as an alternative to the traditional acute toxicity methods, with the aim of reducing both the numbers of animals and the level of pain associated with acute toxicity testing. The stimuli for the development of the FDP were a combination of ethical and scientific concerns regarding the traditional methods that use lethality as the key endpoint.

Evident toxicity is a general term describing clear signs of toxicity following administration of test substance, such that an increase to the next highest fixed dose would result in the development of severe toxic signs and probably mortality.

Underpinning the FDP is a belief that the toxic profile of a substance can be characterized with sufficient reliability for most regulatory situations without the need for the identification of a lethal dose. That is, observations made at non-lethal doses will allow substances to be ranked, or classified, according to their acute toxicity, provide information to aid dose level selection for repeat dose studies and provide hazard data for use in a risk assessment.

Fixed dose levels of 5, 50 and 500 mg/kg were initially chosen as dose levels that would be expected to allow the identification of a dose producing evident toxicity for the majority of substances. These doses also provide information that lead to a similar classification to that based on the LD50 value. The assumption that the severe toxicity/mortality will result at the next highest fixed dose from that producing evident toxicity was a pragmatic one, based on general experience. The validity of this assumption was tested in the subsequent extensive validation exercises that provided a comparison between classification (EU system) resulting from the FDP and that based on the LD50 value obtained from guideline 401.

The test is a group sequential procedure and uses five animals of each sex at each dose. Four preassigned starting levels are possible.

As a preliminary validation step, a literature-based survey of acute toxicity data on 153 substances was conducted, which suggested that for about 80% of these substances classification using the FDP would be the same as that based on the LD50 value. About 14% of the substances would probably be classified in a less severe category and the remainder could be classified in a more severe category (2). The results of a national validation study involving 5 laboratories and 41 substances were published in 1987 (3) followed by an international validation study involving 33 laboratories in 11 countries and 20 substances, published in 1990 (4). The validation studies showed that even with the use of fewer animals and the use of evident toxicity as an endpoint there were no significant inter-laboratory variations in the test results. In relation to classification, the FDP was in agreement with 401 for about 80% of tests, produced a less severe classification in about 16% of tests and a more severe classification in about 3% of tests.

During the validation procedure, a fixed dose of 2000 mg/kg was added to provide more information on substances of low acute toxicity. Also, a sighting study was added as an integral part of the method, to assist the selection of an appropriate starting dose and to provide additional information on the acute toxicity profile of the substance if the sighting study is carried to it completion.

The FDP was published as an OECD Test Guideline in 1992. The performance of the FDP was subjected to biometric analysis in 1992 (5) and 1995 (6). The likelihood of the FDP producing the same classification (EU system) as that based on the LD50 value was estimated for a range of slopes and LD50 values. The mathematical model predicted that for substances with a dose-response slope for lethality of less than about 2, the FDP was likely to lead to a more severe classification that guideline 401. If the slope was between 2 and 6, the FDP was most likely to lead to the same classification. However, for substances with a slope of more than about 6, there was an increasing likelihood of less severe classification; for example, assuming an LD50 of 75 mg/kg and a slope of 6, the FDP classification is more likely to be in the harmful category than the correct toxic category.

B. Point estimate of LD50: The FDP was not originally designed to determine a point estimate of LD50. However, a rule of thumb was developed that permits an approximate LD50 range to be inferred from the classification that results from an FDP. The ability of the FDP to correctly classify (i.e. assign to an LD50 range) in comparison with methods in which the LD50 is estimated is discussed above.

C. Confidence limits on the estimate of LD50: Since the FDP was not designed to determine a point estimate of LD50, confidence limits are also not estimated.

D. Estimate of the slope of the dose-response curve for lethality: The dose-response slope cannot be estimated using the FDP, although some information on dose-response relationship may be available from a sighting study and when more than one fixed dose is used in the main study.

E. Confidence limits on the slope of the dose-response curve for lethality: Confidence limits on the dose-response slope are not provided by the FDP.

F. Dose-effect curve for the LD50: Since lethality is not the preferred endpoint for the FDP, toxicological effects seen only at dose levels close to a lethal dose will not be observed. However, it has been shown in a number of validation and comparative studies (2,3,4,7,8) that while there were a number of instances where clinical signs observed in FDP tests differed from those observed in 401 tests, in only a few cases were these meaningful. In the majority of cases, the clinical signs observed in 401 tests and not observed in the FDP tests were non-specific signs of approaching death.

G. Point estimate of an LD10: The ability of the FDP to predict the LD10 has not been assessed. However, biometric analysis indicated that the most likely classification resulting from the FDP depends on the LD7 of the substance (6), suggesting that this procedure can reliably produce a point estimate of the LD7.

Acute Toxic Class, Guideline 423

A. Principles underlying the test method: The acute toxic class (ATC) method has been developed for hazard assessment, for hazard classification purposes, and for risk assessment. The

method enables the toxicologist to allocate chemical substances to all classification systems currently in use (Example: the LD50 is between 50 and 500 mg/kg body weight) (9,13). It is a group sequential procedure using three animals of one sex per step. Three preidentified starting doses are possible. Three animals of the opposite sex are then dosed at the final dose level used with the first sex. The method was tested in validation studies with animals. Very good congruent results were obtained with animal data and biometrical evaluations, being in the range of 88% (9-13).

The ATC Method is based on the probit model; i.e., the dose-response relationship follows the Gaussian distribution for log-dose values with two parameters, the mean (LD50) and the slope ß in probit units based on the log-scaled dose-axis (logarithm according to base 10). Then, following the test scheme of the method, expected probabilities of a correct, of a lower and of a more stringent classification in dependence on the true oral LD50 value of a substance and its slope can be derived. Also expected numbers of animals used and of moribund/dead animals can be calculated.

The classification procedures were developed in such a manner that on the one hand the probabilities of correct classification are large, and on the other hand the test procedures are simple enough for practical use.

The test doses have been selected with respect to the classification system of chemicals and liquid pesticides of the European Union. It has been shown that

- in the case when test doses and class limits are identical in general the probabilities of correct classification are greater than otherwise.
- the minimal distance factor between two neighboring toxic classes has to be 4 for slopes of ß≥1 to achieve a probability of correct classification of at least 0.5 for at least one LD50 value in each class.
- for a slope of ß≥1 the probability of an allocation to a lower than correct toxic class is limited to 0.256.
- the expected numbers of animals are on average 30% compared to the Guideline 401 (1981) or 45% according to Guideline 401 (1987).
- sex differences with respect to classification are addressed by classifying the substance according to its acute toxicity to the more sensitive sex.
- the classification procedure can be further refined by carrying out a second option - taking into consideration additional class limits as for example 50 or 500 mg/kg body weight.
- this method can be carried out for all acute oral classification systems currently in use.
- there is only a low dependence on the starting dose with respect to classification results, especially for slopes of ß>1. With increasing slopes or increasing LD50 values this influence decreases and tends toward zero for an unlimited increase of ß or LD50. Also for infinitely low values of LD50 the influence becomes zero.
- there is a strong dependence on the starting dose with respect to expected numbers of animals used and of moribund/dead animals. Therefore an appropriate starting dose should be near the true LD50 of the substance to be tested, which leads on average to the least number of animals used.

B. Point estimate of LD50: The ATC was not designed to determine a point estimate of LD50. However, a point estimate of the LD50 can be calculated by the maximum likelihood method providing there are at least two doses with mortality rates not equal to 0% or 100%. However, the probability of this case is rather low because the distance between two neighboring doses is 8- to 10-fold and no more than six animals per dose are used (12).

C. Confidence limits on the estimate of LD50: The ATC was not designed to determine a point estimate of LD50. However, confidence limits on the LD50 can be calculated by the maximum likelihood method providing there are at least three doses, two of which must have mortality rates not equal to 0% or 100%.

D. Estimate of the slope of the dose-response curve for lethality: The ATC was not designed to determine the slope of a dose-response curve for lethality. However, an estimate of the slope of the dose-response curve can be calculated by the maximum likelihood method providing there are at least three doses, two of which must have mortality rates not equal to 0% or 100%.

E. Confidence limits on the slope of the dose-response curve for lethality: Confidence limits on the dose-response slope are not provided by the ATC. However, confidence limits on the slope can be calculated by the maximum likelihood method providing there are at least three doses, two of which show the selected effect and are not equal to 0% or 100%.

F. Dose-effect curve for the LD50: The ATC was not designed to determine a dose-effect curve for the LD50. However, dose-effect curves can be calculated by the maximum likelihood method providing there are at least three doses, two with the specific toxic signs not present in 0% or 100% of the animals.

G. Point estimate of an LD10: The ATC was not designed to determine a point estimate of LD10. However, a point estimate of the LD10 can be calculated by the maximum likelihood method providing there are at least two doses with different mortality rates not equal to 0% or to 100%.

Up-and-Down Method, Guideline 425

A. Principles underlying the test method: The concept of the up-and-down (UDP) testing approach (sometimes called a Staircase Design) was first described by Dixon and Mood (14,15). There have been papers on such issues as its use with small samples (16) and its use with multiple animals per dose (17). One of the most extensive discussions appears in a draft monograph prepared by W. Dixon and Dixon Statistical Associates for a U.S. National Institutes of Health [NIH] Phase I Final Report, Reduction in Vertebrate Animal Use in Research, produced under SBIR Grant No. 1-R43-RR06151-01(18). This draft monograph is available from its author for a fee or from the National Center for Research Resources of the NIH to individuals under the Freedom of Information Act.

In 1985, Bruce proposed the use of the UDP for the determination of acute toxicity of chemicals (19). While there exist several variations of the up-and-down experimental design, Guideline 425 is based on the procedure of Bruce as adopted by ASTM in 1987 (20). The UDP calls for

dosing individual animals of a single sex, usually females, in sequence at 24-hour intervals, with the initial dose set at "the toxicologist's best estimate of the LD50." Following each death (or moribund state) the dose is lowered; following each survival, it is increased, according to a prespecified dose progression factor. If a death follows an initial direction of increasing doses, or a survival follows an initial direction of decreasing dose, four additional animals are tested following the same dose adjustment pattern and then testing is ended. The OECD 425 protocol calls for a default dose progression factor of 1.3 and default σ for maximum likelihood calculations of 0.12 (i.e., log(1.3)).

The first animal is dosed at the toxicologist's best estimate of the LD50. When there is no information on the substance to be tested, for animal welfare reasons it is recommended in the guideline to use the starting dose of 200 to 500 mg/kg body weight.

B. Point estimate of the LD50: From the data a point estimate of the LD50 is calculated using the maximum likelihood method (21,22), provided a suitable historical or other sound estimate of the standard deviation can be employed.

C. Confidence limits on the estimate of LD50: From the data confidence limits around the LD50 value can be calculated using the maximum likelihood method (21,22), provided a suitable historical or other sound estimate of the standard deviation can be employed. However, built into the calculation is a presumption that the parameter σ (standard deviation) is known. σ is the reciprocal of the slope of the probit versus log 10 dose line. An estimate of σ of 0.12 is used unless a better generic or case-specific value is available. The method indicates that the σ value for a previously tested related substance can be used. For compounds of high toxicity with steep slope, this assumption will have little effect on the estimate of the LD50, but the standard error of that estimate is affected and may be unreliable (23).

D. Estimate of the slope of the dose-response curve for lethality: Some dose response information will usually be gained if more than one dose level is used, but an accurate dose response cannot be determined with the procedure as written since default assumptions usually place the σ at 0.12. Dixon (18) has proposed methods to improve the accuracy of the dose-response curve. These require increased numbers of animals but usually less than the guideline 401. These methods are not described in the current OECD protocol.

E. Confidence limits on the slope of the dose-response curve for lethality: Dixon (18) has proposed methods to improve the accuracy of the dose response estimate including determining the confidence limits on the slope of the dose-response curve. These require increased numbers of animals but usually less than guideline 401. These methods are not described in the current OECD protocol.

F. Dose-effect curve for the LD50: Some dose effect information will usually be gained if more than one dose level is used, but an accurate dose effect cannot be determined with the procedure as written since typically some doses will have only one observation. Dixon (18) has proposed methods to improve the accuracy of the dose response estimate. These would also improve a dose-effect estimate but require increased numbers of animals but usually less than guideline 401. These methods are not described in the current OECD protocol.

G. Point estimate of an LD10: The UDP as described in Guideline 425 does not estimate an LD10. Dixon (18) discusses the use of a staircase approach to the estimation of percentage points other than LD50. Such an approach could be explored when LD10 estimates are needed.

References

1. Finney, D J (1971) Probit Analysis, Cambridge University Press, Cambridge, U.K.

2. British Toxicology Society (1984) Special report: a new approach to the classification of substances and preparations on the basis of their acute toxicity, Human Toxicol., 3:85-92.

3. Van den Heuval, M J, A D Dayan and R O Shillaker (1987). Evaluation of the BTS approach to the testing of substances and preparations for their acute toxicity,Human Toxicol., 6:279-291.

4. Van den Heuvel, M J, D G Clark, R J Fielder, P P Koundakjian, G J A Oliver, D Pelling, N J Tomlinson, and A P Walker (1990). The international validation of a fixed-dose procedure as an alternative to the classical LD50 test, Fd. Chem. Toxicol., 28:469-482.

5. Whitehead, A and R N Curnow (1992). Statistical evaluation of the fixed-dose procedure, Fd. Chem. Toxic., 30:313-324.

6. Stallard N and A Whitehead (1995). Reducing numbers in the fixed-dose procedure, Human Expt. Toxicol., 14:315-323.

7. Lipnick, R L, J A Cotruvo, R N Hill, R D Bruce, K A Stitzel, A P Walker, I Chu, M Goddard, L Segal, J A Springer, and R C Myers (1995). *Comparison of the Up-and-Down, Conventional LD50, and Fixed-Dose Acute Toxicity Procedures*. Fd Chem. Toxic. 33: 223-231.

8. Yam, J, P J Reer, and R D Bruce (1991). *Comparison of the Up-and-Down Method and the Fixed-Dose Procedure for Acute Oral Toxicity Testing*. Fd Chem. Toxic. 29:259-263.

9. Diener, W; U Mischke, E Schlede and D Kayser (1995). *The biometric evaluation of the OECD modified version of the acute toxic class method (oral).* Arch.Toxicol. 69: 729-734.

10. Diener, W; and E Schlede. (1996). *Brief an den Herausgeber: ML Prinzip und ATC-Methode*. ALTEX 13(4): 238-239.

11. Diener, W; and E Schlede (1996). *Letter to the Editor: FDP and ATC method: a mathematical comparison.* Human Experim.Toxicol. 15: 855-856.

12. Diener, W; L Siccha, U Mischke, D Kayser and E Schlede (1994*). The biometric evaluation of the acute-toxic-class method (oral).* Arch.Toxicol. 68: 599-610.

13. Schlede, E; U Mischke, W Diener and D Kayser. (1995). *The international validation study of the acute-toxic-class method (oral).* Arch.Toxicol. 69: 659-670.

14. Dixon, W J, and A M Mood (1948). A method for obtaining and analyzing sensitivity data. J. Amer. Statist. Assoc. 43:109-126.

15. Dixon, W J (1991). Staircase Bioassay: The Up-and-Down Method. Neurosci. Biobehav. Rev. 15:47-50.

16. Brownlee, K A , J L Hodges, Jr., and M Rosenblatt (1953). J Amer. Stat. Assoc., 48:262-277.

17. Hsi, B P (1969). J Amer. Stat. Assoc., 64:147-162.

18. Dixon, W J and Dixon Statistical Associates (1991). Design and Analysis of Quantal Dose-Response Experiments (with Emphasis on Staircase Designs).

19. Bruce, R D (1985). An up-and-down procedure for acute toxicity testing. Fundam. Appl. Tox., 5:151-157.

20. ASTM (1987) E 1163-87, Standard test method for estimating acute oral toxicity in rats. American Society for Testing Materials, Philadelphia PA, USA.

21. Dixon, W J (1965). The up-and down method for small samples. J. Amer. Statist. Assoc., 60:967-978.

22. Finney, D J (1971). Probit Analysis, 3rd ed., Cambridge University Press, Cambridge, England, 50-80.

23. Stallard, N and A Whitehead (1996). A preliminary statistical evaluation of the up-and-down procedure. Medical and Pharmaceutical Research Unit, University of Reading.