revcia.
6/06/01

Description of Performance and Confidence Intervals for the Revised Up-and-Down Procedure (UDP) for Acute Oral Toxicity

June 6, 2001

Prepared by:
The UDP Technical Task Force
U.S. Environmental Protection Agency

Submitted to:
The Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM)

Description of Performance and Confidence Intervals for the Revised Up-and-Down Procedure for Acute Oral Toxicity

Table of Contents

List of Tables

Appendix A

Appendix B

List of Figures

Figure No. **Page**

Appendix A

.

List of Abbreviations and Acronyms

AEGL	Acute Exposure Guideline Level
ASTM	American Society for Testing of Materials
ATWG	Acute Toxicity Working Group
CFR	Code of Federal Regulations
CPSC	Consumer Product Safety Commission
CRP	Child Resistant Packaging
EPA	Environmental Protection Agency
FHSA	Federal Hazardous Substances Act
FIFRA	Federal Insecticide, Fungicide and Rodenticide Act
HAZMAT	Hazardous Materials
ICCVAM	Interagency Coordinating Committee on the Validation of Alternative Methods
MLE	Maximum Likelihood Effect
OECD	Organisation for Economic Co-operation and Development
PM	Performance Measure
PPPA	Poison Prevention Packaging Act
QA/QC	Quality Assurance/Quality Control
TG	Test Guideline
UDP	Up-and-Down Procedure

Executive Summary

The draft Revised Up-and-Down Procedure guideline recommends profile likelihood methods, using established theory, for most instances where confidence intervals can be obtained. These are widely used methods that take into account uncertainty in the mean of the population from which the data are drawn. While other types of intervals could have been developed (e.g., bootstrap, isotonic, Bayesian), profile likelihood methods are often used for their practicality and were readily available when the originally proposed Up-and-Down Procedure supplemental test for slope and confidence interval was deleted.

Data gathered under the Revised Up-and-Down Procedure fall into one of five scenarios. Simulations are provided for the performance of the Revised Up-and-Down Procedure in these five cases. Simulations and the fundamental mathematical structure have indicated that in three of these scenarios, standard probit procedures cannot be applied with data generated using the Revised Up-and-Down Procedure. (This can also happen with other multi-treatment-level designs.) Therefore, special statistical procedures are proposed for use in these cases. The point estimates are specified in the test guideline. These circumstances also define availability of the profile likelihood confidence interval and special procedures are proposed for interval estimation.

Calculation of the profile likelihood requires maximizing the likelihood function while holding the term for the LD50 at a fixed assumed value. At each fixed assumed LD50, the likelihood will be maximized by some particular value of the slope. Calculation of the profile likelihood confidence intervals requires calculating the profile likelihood for different values of fixed assumed LD50s with their corresponding profile maximizing slopes and finding the value for which the profile likelihood equals a critical value. This is a computationally-intensive procedure. Consequently, special-purpose software has been developed.

Each of the methods considered can be applied in some scenarios but not in others. In a small percentage of cases no confidence interval would be provided.

1.0 Performance and Confidence Intervals for the Revised Up-and-Down Procedure for Acute Oral Toxicity

1.1 Background and History

Calculation of confidence intervals gives the user a basis for evaluating how to incorporate test results into regulatory applications. Therefore, a confidence interval calculation was included in previous versions of the Up-and-Down Procedure (UDP) guideline (both OECD 1998 and ASTM 1998 and prior). Following deletion of the proposed supplemental procedure from the previous draft Revised UDP, another method was needed to assist the investigator using the UDP to calculate a confidence interval.

The statistical procedure in the previous version of OECD Test Guideline 425 did not produce a true confidence interval because it relied on an assumed value of *sigma* (the slope parameter). This limitation was pointed out in Bruce (1985) and by the ICCVAM UDP Peer Panel (July 2000). While the calculation of the LD50 estimate proposed for the Revised UDP also uses an assumed *sigma*, a separate statistical procedure is proposed for obtaining the confidence intervals for the data. This confidence interval procedure does not rely on the assumed value of *sigma*.

A provision for confidence interval calculation has been added to the statistical analysis of the LD50 estimate from the Up-and- Down Procedure (UDP). Information on the quality of a point estimate and the data from which it is derived are important in understanding the outcome of the test. A confidence interval can be viewed as providing plausible bounds on the value of the LD50 based on the data collected in the particular study. A description of the added feature for calculation of confidence intervals has been inserted at paragraph 40 in the latest revision of the UDP guideline.

An OECD expert group agreed with the addition of the feature for calculation of confidence intervals. Subsequently, the Acute Toxicity Working Group (ATWG) decided to bring the confidence interval insertion to the UDP Peer Panel for comment. Pursuant to these events, a government contract for software development was initiated. The software package for the main test provides (a) information to the experimenter on how many animals are to be dosed and (b) the statistical procedure for estimating the LD50 and confidence interval. A plan for verification of the software package is included in Section 3.0 of this document.

1.2 Regulatory Applications of Confidence Intervals

Statisticians distinguish between point and interval estimation of parameters. Point estimation results in a single value estimate for a parameter, as provided, for example, by the UDP procedure for estimating the LD50. Interval estimation is expressed in a lower and upper bound for an interval that has a known probability of containing the true value of the parameter. That probability is called the confidence coefficient.

To compute a confidence interval, a statistical algorithm needs both the desired confidence coefficient and the experimental data. In the case of the UDP, the experimental data are the doses and responses. The statistical algorithm is designed to compute a 95% confidence interval, which is the typical confidence coefficient in statistical practice. However, the algorithm is not exact but approximate, so that in some situations, the interval will not provide the desired coverage or may provide more than the desired coverage. The results from simulation studies in Appendices A and B of this document will be useful for experimenters to assess if the data and estimated LD50 are producing confidence intervals that are in the same range as simulated intervals that have the desired coverage.

At a given confidence coefficient, the width of the confidence interval is a result of the underlying variability in the dose-response curve. Wider intervals imply less precision in the estimate of the LD50, and also that replications of this experiment with the same compound and animal species under identical conditions could produce meaningfully different LD50 estimates. Moreover, in comparing two different chemical compounds, the widths and locations of the associated confidence intervals provide an indication as to whether the data used to estimate the LD50s lead to estimates precise enough to consider one chemical's LD50 larger or smaller than the other.

Confidence intervals, provided they can be calculated, describe the range of estimates that are consistent with the data seen. In addition, when comparisons of compounds are made using estimated LD50s, confidence intervals give a sense of the robustness of the comparisons. Consequently, any confidence interval is seen as adding descriptively to the data at hand and is not used to exclude information.

Weight-of-evidence deliberations for risk assessments already rely on confidence intervals together with other study details and results. Hazard identification also relies on confidence intervals to assess the meaning of lethality estimates. Such regulatory determinations include:

! decisions about special packaging requirements for products to which children might be exposed,
! registration and reregistration of pesticides,
! review of potential hazard or risk of chemicals to endangered species, and
! hazard identification for consumer and industrial chemicals and mixtures.

Other regulatory instances where confidence intervals are reported include assignment of chemicals or mixtures to toxicity categories used in the regulation of workplace or consumer products, as well as in:

- development of Acute Exposure Guideline Levels (AEGLs; any of three ceiling airborne exposure values for the general public applicable to emergency exposure periods ranging from less than one hour to eight hours);

- routine decisions about child-resistant packaging and labeling;

- classification of substances (e.g., pesticide active ingredients-technical grade);

- for determining hazardous materials (HAZMAT) categories in transport;

- classification of industrial chemicals used in the workplace; and

- classification of mixtures such as pesticide and end-use products (the formulated product).

1.3 Examples of Regulatory Applications of Confidence Intervals

1.3.1 U.S. Consumer Product Safety Commission

Application of Confidence Interval in Evaluation of Hazard and Risk

The confidence interval is important for appropriate evaluation and use of acute toxicity data. An LD50 with a narrow confidence interval that falls within a classification class criteria can be used reliably, whereas an LD50 with a very wide confidence interval (2 mg/kg to 5000 mg/kg) spanning multiple class criteria has to be used very judiciously. The use of numerical values of the LD50 estimate along with the calculated confidence interval becomes more important in a risk assessment (likelihood of injury/illness determination) or when the toxicities of two substances are compared.

The confidence interval is an integral part of a statistical evaluation of toxicity data and its use will be increasingly more important since the number of animals used in testing is being decreased for animal welfare reasons. The number of animals used in a test impacts the size of the confidence interval. Generally, when fewer animals are used, the confidence interval is wider. The width of the confidence interval would determine appropriate use of the data for classification purposes, in risk assessment, or for comparison of toxic potential of two substances, etc.

Regulatory Citations for Acute Toxicity Data including Confidence Intervals:

For a substance to be defined as “hazardous substance”, the Consumer Product Safety Commission under its Federal Hazardous Substances Act (FHSA, 16 CFR 1500.3) requires a two-part determination: 1) that a substance/product has a toxic property, and 2) that it may cause substantial personal injury or substantial illness during or as a proximate result of any customary or reasonably foreseeable handling or use, including reasonably foreseeable ingestion by children. The toxicity data should be statistically significant and shall be in conformity with good pharmacological practices. A toxicity numerical value such as an LD50 should be accompanied by an index of variability such as a confidence interval.

The Commission also enforces the Poison Prevention Packaging Act (PPPA). The PPPA regulations for exemptions (16 CFR 1700.9 (a)(4)) state:

> “(4) In view of the fact that LD50 values in themselves do not necessarily reflect a true estimate of the overall toxic potential of a substance, LD50 determinations should, where an LD50 value may be calculated, include:
>
> (I) The LD50 value with 95 percent confidence limits; (ii) a slope determination for the dose response curve, including 95 percent confidence limits; and (iii) a description of the statistical method employed in the analysis of such data (with proper citation) as well as the statistical analysis itself.”

1.3.2 U.S. Environmental Protection Agency (EPA)

<u>Regulatory Citations for Pesticides under Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA):</u>

40 CFR 158.80 sets forth general policy for acceptability of data as follows:

> "In evaluating experimental design, the Agency will consider whether generally accepted methods were used, sufficient numbers of measurements were made to achieve statistical reliability, and sufficient controls were built into all phases of the experiment. The Agency will evaluate the conduct of each experiment in terms of whether the study was conducted in conformance with the design, good laboratory practices were observed, and results were reproducible."

At 40 CFR 158.202(e)(1) for human health:

> "Determination of acute oral, dermal and inhalation toxicity is usually the initial step in the assessment and evaluation of the toxic potential of a pesticide. These data provide information on health hazards likely to arise soon after, and as a result of short term exposure. Data from acute studies serve as a basis for classification and precautionary labeling."

At 40 CFR 158.202 (h)(2) for nontarget organisms in the environment:

> "The short term acute laboratory studies ... are used to establish acute toxicity levels of the active ingredient to the test organisms; to compare toxicity information with

> measured or estimated pesticide residues in the environment in order to assess potential impacts on fish, wildlife and other nontarget organisms; and to indicate whether further laboratory and/or field studies are needed."

Hazard Classification and Risk Assessment of Pesticide Formulations for Human Health:

40 CFR 156.10 provides for hazard labeling of pesticides; Part 152.160 provides for classification of pesticides; and Parts 152.170, 152.171, and 152.175 provide for restricted use of pesticides. Historically, Agency reviewers have tended to consider only the LD50 value in assigning a pesticide formulation to a toxicity category in terms of its oral or dermal toxicity. The traditional acute toxicity study could be relied upon to provide relatively manageable confidence intervals. Confidence limits associated with the LD50 values have generally been reported by the performing laboratories. They are usually included in Agency review summaries.

This situation has changed. With the use of acute toxicity testing protocols that minimize the numbers of animals tested, it becomes more important for Agency toxicologists to consider not only the findings of a study, but also its inherent statistical limitations, in any interpretation and regulatory decision. As a result, in a situation where an LD50 estimate falls so close to a classification boundary that the confidence limits (or bracketing range) include values well below the boundary value, Agency reviewers must take a conservative approach, and classify the test material in the more toxic category. Under these circumstances, the toxicology reviewers would normally feel comfortable with the use of 90% confidence limits, as there would then be only a 5% probability that the LD50 value would be below the lowest value of the confidence interval range. However, they would also have to take into consideration the presence or absence of symptoms of toxicity in the test animals, particularly in situations when severe and/or life-threatening reactions occur at lower dose levels with subsequent recovery and no mortality.

FIFRA Section 25(c)(3) authorizes the Agency to establish Child-Resistant Packaging (CRP) standards, consistent with those under the authority of the Poison Prevention Packaging Act (Public Law 91-601), to protect children from serious injury or illness resulting from accidental ingestion or contact with pesticides. CRP is required for residential use products with an LD50 value of 1500 mg/kg and less, or meeting any of the other toxicity criteria in 40 CFR 157.22(a). If there is a $5% probability that the oral LD50 value is at or lower than 1500 mg/kg, then a toxicology reviewer would recommend the use of CRP. Taking into consideration the emphasis on protecting children from serious injury or illness, an Agency toxicologist would also evaluate the occurrence and severity of toxicological symptoms in an acute oral LD50 study at doses below which mortality occurs.

Environmental Assessment of Pesticides:

Confidence intervals are used in risk assessment for the same purpose as in general statistics to express the "level of confidence" that a sample mean (or other summary statistic) represents the true population mean. Toxicity tests performed for regulatory purposes typically are limited in several ways (i.e., sample size, standardized laboratory conditions, etc.). For these reasons, a sample mean (or statistic such as LD50) is generally only a very rough estimate of the actual population being sampled in the test. The confidence interval in this case does describe the level of confidence in the true value, but also serves the reader as a measure of the utility of the test overall. Confidence intervals support compliance with Agency Quality Assurance/Quality Control (QA/QC) principles of precision. Confidence intervals are principally a data QA/QC measure. Point estimates should not to be reported without some measure of precision. Moreover the Agency's QA/QC policies state that the Agency is to use data of known precision. In rating a test result submitted for registration or re-registration of a pesticide, the confidence interval can be considered along with other measures of the validity of the test such as availability of dose response of the test population's tolerance to the pesticide.

Traditionally, toxicity tests for nontarget species are designed to address "dose response" and a narrow confidence interval is an indication of how well a "dose response" was achieved in the study. If the precision of an obtained LD50 study is inadequate, the Agency needs to know that. A good understanding of "dose response" is also useful in risk assessment for extrapolating effects across species and establishing distributional bounds for probabilistic assessments.

The Agency plans to develop methods for probabilistic risk assessments for pesticides which will use confidence intervals from acute tests of nontarget species to describe uncertainty. The uncertainty in the LD50 estimate is an important component in estimating the overall uncertainty in a probabilistic risk assessment. Confidence intervals are necessary for estimating the overall uncertainty/variability in a distribution of risk.

Endangered Species Assessments for Pesticides:

Confidence intervals for the LD50 value are not directly used in assessing effects on endangered species because the intent for endangered species is to protect individuals and not simply the typical representative (i.e., at the population mean). The slope allows the reviewer to determine any mitigation provisions needed to attain an endangered species no-effect level, which is what is necessary under the Endangered Species Act. No-effect levels, such as can be obtained by using the slope in conjunction with the LD50, are used for this purpose. Absent a reliable estimate of the no-effect level, a safety factor is applied to the LD50 value, and the reliability of the LD50 value, as indicated by the confidence intervals is an important feature of the test results.

Setting Acute Exposure Guideline Levels under the Superfund Amendment and Reauthorization Act (SARA):

Acute Exposure Guideline Level-3 (AEGL-3, one of three ceiling airborne exposure values for the general public applicable to emergency exposure periods ranging from less than one hour to eight hours) is the airborne concentration (expressed as ppm and mg/m^3) of a substance at or above which it is predicted that the general population, including "susceptible" but excluding "hypersusceptible" individuals, could experience life-threatening effects or death. Airborne concentrations below AEGL-3 but at or above AEGL-2 represent exposure levels which may cause irreversible or other serious, long-lasting effects or impaired ability to escape.

When a confidence interval is available for an LD50, it may be used to discriminate between studies for use in development of an AEGL-3, to decide whether a study can be used for calculating the LC01 that is the basis for an AEGL-3, or to determine the uncertainty factor in calculation.

U.S. EPA’s Policy for Risk Characterization:

The U.S. EPA's Science Policy Council recently issued a Risk Characterization Handbook (EPA 100-B-00-002, Dec. 2000). It focuses on how to integrate "information from the ... components of the risk assessment and [synthesize] an overall conclusion about risk that is complete, informative, and useful for decision makers." Here are some excerpts:

(p. 11) "The overall risk characterization lets the manager, and others, know why the U.S. EPA assessed the risk the way it did in terms of the available data and its analysis, uncertainties, alternative analyses, and the choices made. A good risk characterization will restate the scope of the assessment, express results clearly, articulate major assumptions and uncertainties, identify reasonable alternative interpretations, and separate scientific conclusions from policy judgments."

(p. 13) "Risk characterization communicates the key findings and the strengths and weaknesses of the assessment through a conscious and deliberate transparent effort to bring all the important considerations about risk into an integrated analysis by being clear, consistent and reasonable. Remember, though, unless you actually characterize the assessment, the risk assessment is not complete - - risk characterization is an integral component of every risk assessment. As an example, just giving the quantitative risk estimate ('the number') is not a risk characterization."

(p. 21) "Your specific responsibilities [as a Risk Assessor] are to:

...d) Describe the uncertainties inherent in the risk assessment and the default positions used to address these uncertainties or gaps in the assessment

...f) Put this risk assessment into a context with other similar risks that are available to you and describe how the risk estimated for this stressor, agent or site compares to others regulated by EPA"

(p. 36) "[Elements that affect a Risk Characterization include]:

...f) Variability (Section 3.2.7)

g) Uncertainty (Section 3.2.8)..."

(p. 37) "For each stage of the assessment for human health or ecological risks, the assessor identifies:

a) The studies available and how robust they are (e.g., have the findings been repeated in an independent lab)

b) The major risk estimates calculated, the assumptions and the extrapolations made during the estimated risk calculations, and the residual uncertainties and their impact on the range of plausible risk estimates. Your description of the risk estimate should indicate what you are assessing (e.g., individual, population, ecosystem) and include such things as the high end and central tendency estimates.

...f) Variability (see Section 3.2.7)"

(p. 40) "3.2.7 How Do I Address Variability?

The risk assessor should strive to distinguish between variability and uncertainty to the extent possible (see 3.2.8 for a discussion of uncertainty). Variability arises from true heterogeneity in characteristics such as dose-response differences within a population, or differences in contaminant levels in the environment. The values of some variables used in an assessment change with time and space, or across the population whose exposure is being estimated. Assessments should address the resulting variability in doses received by the target population. Individual exposure, dose, and risk can vary widely in a large population. Central tendency and high end individual risk descriptors capture the variability in exposure lifestyles, and other factors that lead to a distribution of risk across a population."

"3.2.8 How Do I Address Uncertainty?

Uncertainty represents lack of knowledge about factors such as adverse effects of contaminant levels which may be reduced with additional study. Generally, risk assessments carry several categories of uncertainty, and each merits consideration. Measurement uncertainty refers to the usual error that accompanies scientific measurements -- standard statistical techniques can often be used to express measurement uncertainty..."

1.4 Calculation of Confidence Intervals for the Revised UDP

Inserted text at paragraph 40 of the Revised UDP states:

> "40. Following the main test and estimated LD50 calculation, it may be possible to compute interval estimates for the LD50 at specified confidence using a profile-likelihood-based computational procedure. Such an interval utilizes information from the doses where accumulated response was neither 0% nor 100% (intermediate doses). Instead of employing an assumed *sigma*, however, the procedure identifies bounds on LD50 estimates from a ratio of likelihood functions optimized over *sigma* (profile likelihoods). Procedures are also included for certain circumstances where no intermediate doses exist (for instance, when testing has proceeded through a wide range of doses with no reversal or where doses are so widely spaced that each animal

> provides a reversal). Implementing this set of procedures requires specialized computation which is either by use of a dedicated program to be available from OECD or developed following technical details available from OECD."

For many or most studies conducted according to the Revised UDP, standard probit calculations will not be able to provide the basis for a confidence interval. Instead, the Revised UDP uses profile likelihood methods based on established theory for most instances where confidence intervals can be obtained. These are widely used methods that take into account uncertainty in the mean of the population from which the data are drawn. While other types of intervals could have been developed (e.g., bootstrap, isotonic, Bayesian), profile likelihood methods are often used for their practicality and were readily available when the originally proposed UDP supplemental test for slope and confidence interval was deleted.

Profile likelihood confidence intervals are based on the same kinds of functions as the point estimate, namely, the likelihood function and ratios of that function. In addition, the proposed confidence interval uses the same distributional shape assumptions as the point estimate, while making no numeric assumptions about its parameters (i.e., no value for $sigma$ is assumed). In order to reduce such assumptions, this method is computationally intensive using modern methods. Consequently, a specialized program is needed for its implementation. Software will be provided to users on request or through a web site (e.g., OECD's). The OECD Expert Meeting in August 2000 supported this proposal.

The calculation should and does take advantage of established theory, modern computational methods, and previously used and tested algorithms (Rao, 1973; Bickel and Doksum, 1977; Crump and Howe, 1985; Meeker and Escobar, 1995) and utilizes knowledge of the full sample of observations. Results from doses where no or all animals respond does contribute some information on the LD50, even when a point estimate cannot be calculated.

The methodology for this confidence interval has also been used (previously used and tested algorithms) with estimates beside the LD50, including the limit on a benchmark dose (used in U.S. EPA health risk assessments).

Because similar intervals behave well in similar situations, the proposed confidence intervals are expected to perform appropriately for the Revised UDP. The term "behaving well" means that the intervals will have at least the stated coverage probability in simulated trials; that is, at least 95% of simulated '95% CIs' include the true LD50 (see Appendix A).

Just as with the point estimate, there are some circumstances where a standard approach will have computational problems. For example, as outlined in OECD TG 425 paragraph 42 or Revised UDP paragraph 37; there may be only increasing or only decreasing doses throughout the test. Certain solution choices are suggested and included in the special software.

1.5 Performance Characteristics of the Revised UDP Including Case Examples

Five scenarios or cases can be distinguished for the purpose of describing the performance of the Revised UDP as shown in Table 1. Cases 2 and 4 permit estimation of the LD50 and confidence intervals. Cases 1, 3, and 5 do not permit calculation of either an LD50 using the main method, or a confidence interval using the profile likelihood method. Some response patterns for these cases do provide some information about the location of the LD50. More detail on these cases is below.

Case 2 is the standard two parameter probit estimation situation. The case has intermediate response fractions (at least one animal and less than all animals respond) at some dose that is less than a dose where there was no response. Typically, intermediate response fractions will occur at more than one dose. Point estimates and confidence intervals are available.

Case 4 has a single intermediate response fraction occurring between doses that have no response and doses with complete response. The LD50 can be estimated and confidence intervals can be calculated for this case.

Case 1 has three possible response patterns: (a) all animals responded, (b) no animals responded, or (c) the geometric mean dose is lower for animals that responded than for animals that did not respond. Case 1a suggests that the LD50 is likely to be lower than the lowest dose while Case 1b suggests that the LD50 is likely to be greater than the highest dose. Case 1c suggests a reverse dose-response curve, that is fewer responses occur at higher doses. These inferences can be guaranteed to be true, because response is a chance event.

Case 3 has no intermediate response fractions. At some doses, all animals will respond while at lower doses, no animals will respond. This implies that the LD50 is between highest dose with no response and the lowest dose where complete response. Any value between the two doses is a valid estimate for the LD50. No confidence interval can be computed. The situation is likely to emerge from a chemical with a very steep dose-response curve.

There are two possible situations for Case 5. One possibility has an intermediate response fraction at the highest tested dose and no responses at lower doses. This suggests that the LD50 is around the highest tested dose or possibly greater. The second situation has partial response at the lowest tested dose and complete response at higher doses. Here, the LD50 is likely to be at or below the lowest tested dose. For Case 5 data (as for Case 4 data), the LD50 estimate of the software will be the dose with partial response. The confidence interval will be calculated using profile likelihood.

As noted above, data gathered using the Revised UDP fall into one of five types of summary configurations. Simulations and the fundamental mathematics structure have indicated that in three of these configurations, standard probit procedures (e.g., Finney, 1971) cannot be applied with data generated using the Revised UDP. (This can also happen with other multi-treatment-level designs.) Therefore, special statistical procedures are proposed for use in these cases with the Revised UDP.

The point estimates are specified in the Revised UDP. These circumstances also define availability of the profile likelihood confidence interval and special procedures are proposed for interval estimation.

Calculation of the profile likelihood requires maximizing the likelihood (function) while holding the term for the LD50 at a fixed assumed value. At each fixed assumed LD50, the likelihood will be maximized by some particular value of the slope. Calculation of the profile likelihood confidence intervals requires calculating the profile likelihood for different values of fixed assumed LD50s with their corresponding profile maximizing slopes and finding the value for which the profile likelihood equals a critical value. This is a computationally-intensive procedure. Consequently, these will be incorporated into the special-purpose software under development.

Each of the methods considered can be applied in some cases but not in others. In a small percentage of cases, no confidence interval would be provided.

These cases are outlined in Table 1 and Figures 1 and 2.

Table 1. **Outcomes of the Up-and-Down Procedure: Cases and Confidence Intervals.**

Case #	Definition of Case	Approach Proposed	Possible Findings
1	**No positive dose-response association.** There is no variation in response: a) all animals tested in the study responded, or b) none responded, or c) the geometric mean dose is lower for animals that responded than for animals that did not respond.	No confidence interval proposed, inference related to LD50 questionable.	No statistical results. Possible inferences: 1a) LD50 < lowest dose; 1b) LD50 > highest dose; 1c) reverse dose-response curve
2	**Standard 2-parameter probit estimation.** 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 intermediate response fractions, but holds in some other cases as well.)	Profile loglikelihood computations are straightforward.	The LD50 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 an intermediate response fraction.	Lower bound = highest test dose with 0% response. Upper bound = lowest test dose with 100% response.	High confidence that the true LD50 falls between the two bounding doses. Highest dose with 0% response < LD50 < lowest dose with 100% response.
4	**One partial response fraction, first subcase.** Like Case 3 except that an intermediate response fraction 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.	Profile loglikelihood calculations to be extended to this case by special computations.	The LD50 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).	Profile loglikelihood calculations to be extended to this case by special computations	The LD50 is estimated and its confidence interval calculated. Possible inference: the LD50 is near the dose with the intermediate response fraction.

Figure 1. Predicted Percentage of Cases - LD50 equal to 1500 mg/kg.

Predicted percentages of cases

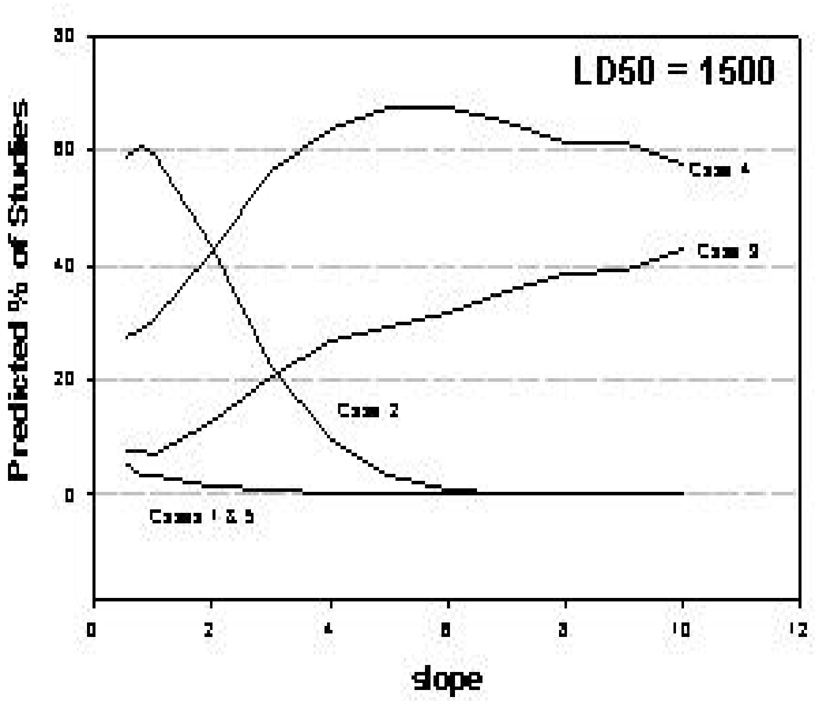
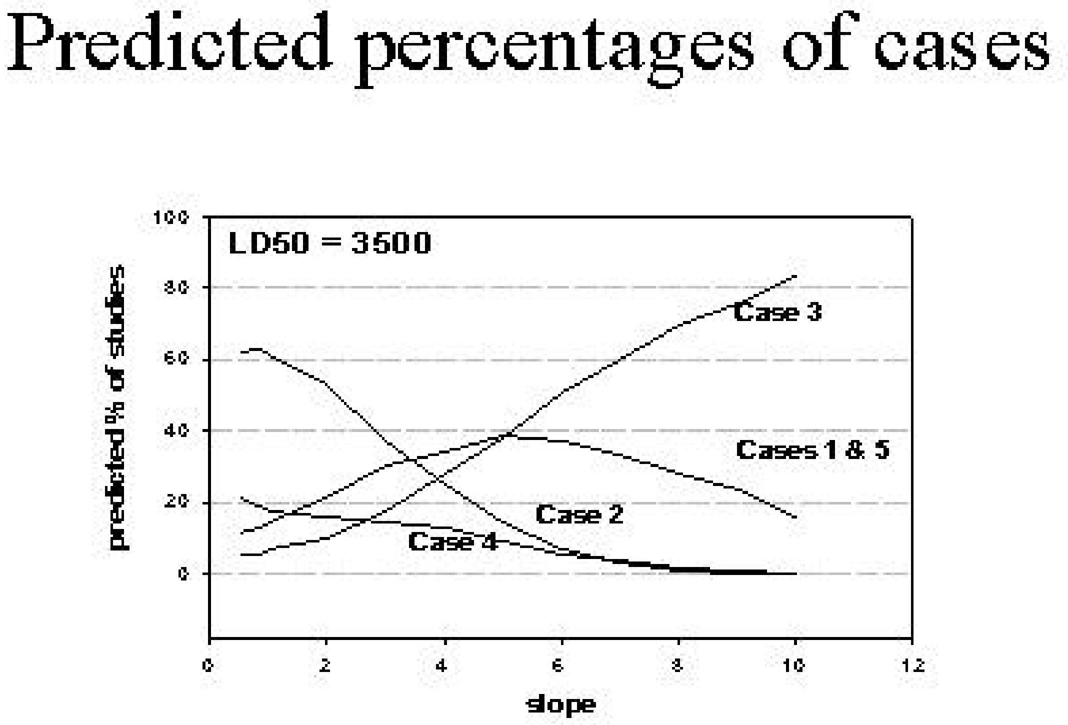


Figure 2. Predicted Percentage of Cases - LD50 equal to 3500 mg/kg.



2.0 LD50 Confidence Bounds for Revised UDP: Statistical Approach and Performance Characterization

2.1 Background.

This section addresses the implementation of confidence bounds for the LD50, for use with acute toxicity data generated in accordance with the Revised UDP. Simulations presented in this document indicate that in a large proportion of cases, standard probit procedures (e.g., Finney, 1971) cannot be applied with data generated using OECD TG 425. Therefore, special statistical procedures are proposed for use with the Up-and-Down Procedure for LD50.

The purpose of this section is to provide an overview of the procedures proposed. Also, simulations are reported to evaluate the performance of the methods proposed. Performance is characterized in terms of the widths of confidence intervals, and in terms of "coverage" probabilities (defined in Section 2.2).

Based on simulations (Section 2.6), it appears that in most cases it will be possible to compute a confidence interval with acceptable performance by one of two methods. In cases where no animals respond at some doses, and all animals respond at some other doses (the latter being greater than the former), the lower bound for the LD50 will be the highest dose associated with no observed responses. Similarly, the upper bound will be the lowest dose associated with response for all animals tested at that dose. In most other cases, it will be possible to compute a bound using the method of profile likelihood (Section 2.4). In particular, it appears that the profile likelihood approach is applicable in most cases where there is only one dose with an intermediate response fraction (neither 0% nor 100% responding), a case that is not handled by standard probit methods. (Proposals for handling various cases are summarized in Section 2.3)

The confidence interval procedures are to be made available in software developed for support of the Revised UDP. The software will also provide point estimates of the LD50 as indicated in the Revised UDP and will evaluate stopping criteria.

The remainder of this section assumes a familiarity with standard probit computations as used in toxicology (Finney, 1971), familiarity with basic statistical procedures (although the definition of a confidence interval is reviewed), and familiarity with the use of Monte Carlo simulation to evaluate the performance of statistical procedures.

2.2 Confidence Intervals: Definition and Related Terminology

Approximate 2-sided 95% confidence intervals will be implemented. Two interpretations of such an interval will be offered in this section. The definition that is most standard is that the probability is 0.95

that the true value of the parameter of interest (here, the LD50) lies within the interval. Here, the parameter of interest is viewed as a fixed constant and the bounds (being based on data) are viewed as random (e.g., Sokal and Rohlf, 1981, particularly Section 7.3). In order for this definition to be useful, the probability of 0.95 must hold at least approximately over the possible values of the parameter of interest, even though the value of that parameter is not know in a given situation.

To understand this interpretation, it may be helpful to reflect on how simulations are used to evaluate a confidence interval (see Section 2.5). In fact, it is common to use simulations to illustrate the concept of a confidence interval (e.g., Sokal and Rohlf, 1981, Figure 7.4).

The probability that the upper and lower bound will enclose the true LD50 is defined to be the *coverage* of the interval. If the coverage of a nominal 95% interval is precisely 95%, then the interval is said to be *exact*. In statistical practice, it is common to use confidence intervals that are not exact but approximate. When intervals are approximate, it is sometimes preferred that they be *conservative*, meaning that the coverage exceeds 95%.

A second interpretation can be particularly helpful for understanding the profile likelihood approach proposed here. According to the second interpretation, a confidence interval for a parameter is to be interpreted as the range of values of the parameter that is consistent with (not excluded by) a particular data set. Thus, Cox and Hinkley state (1974, p. 208) that “foremost is the interpretation that ‘such and such parameter values are consistent with the data.’ ” Confidence intervals can be constructed by inverting statistical hypothesis tests, by defining the confidence interval to be the set of parameter values not rejected using the hypothesis test. In particular, the profile likelihood intervals proposed in this document invert a profile likelihood ratio test.

These two approaches are considered to be consistent. A result given in advanced texts is that a confidence interval with desired coverage can be obtained by inversion of a hypothesis test (e.g., Cox and Hinkley, 1974, Section 7.2; Casella and Berger, 1990, Section 9.2; Bickel and Doksum, 2001, Section 4.2).

2.3 Classification of Cases and Methods Proposed for Particular Cases

Each of the methods considered can be applied in some cases but not in others. In a small percentage of cases, no method of computing a confidence interval is proposed. It is proposed that the selection of a method be based on the classification of cases displayed in Table 2. (Development of this scheme has benefitted from discussions with the OECD acute avian statistics group. See Table 2 footnote.) The rationale for the decisions indicated in this table is as follows.

Case 1. With the stopping rules indicated for the Revised UDP, this case appears to be possible only if testing is stopped at a limit dose (based on non-response for three animals tested in sequence at the dose). No methods are proposed here for cases where there is not an observable relationship between dose and response. In some cases, a binomial test may be used to establish that the LD50 is above or below the range of doses tested, but a significant binomial test requires testing of five or more animals at the same dose, and binomial tests use only data from a single test dose. Some procedures that may be applicable in this case have been developed for avian

Table 2. Classification of Data Cases for Purposes of Confidence Interval Computation for Case 5

Case #	Definition of Case	Approach Proposed
1	**No positive dose-response association.** There is no variation in response (all animals tested in the study responded, or none responded), or the geometric mean dose is lower for animals responding than for animals not responding	no confidence interval proposed, inference related to LD50 questionable.
2	**Standard 2-parameter probit estimation.** One or more animals responded at a dose below some other dose where one or more animals did not respond. The conditions defining Case 1 do not hold. (The definition holds if there are two doses with intermediate response fractions, but holds in some other cases as well.)	profile loglikelihood computations are straightforward
3	**No intermediate response fractions.** One or more test doses is associated with 0% responses and one or more test doses is associated with 100% responses (all of the latter being greater than all of the former), and no test doses are associated with an intermediate response fraction.	lower bound = highest test dose with 0% response. upper bound = lowest test dose with 100% responses.
4	**One partial response fraction, first subcase.** Like Case 3, except that an intermediate response fraction is observed at a single test dose. That dose is greater than doses associated with 0% responses and lower than doses associated with 100% responses.	profile loglikelihood calculations to be extended to this case by special computations
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% responses at all other test doses).	profile loglikelihood calculations to be extended to this case by special computations

Acknowledgement. The development of this scheme was helped by conversations with the OECD avian acute statistics group, which has developed a similar classification (report in press). The avian scheme differs in some points.

acute testing (report in press).

Case 2. In cases where standard probit computations can be applied, it appears that application of the profile likelihood (described in Section 2.4) will be straightforward. The profile likelihood approach is already used in this situation in the U.S. EPA benchmark dose software.

It is common to require, as a condition for probit analysis, that there are at least two test doses with partial response fractions (response fractions not 0% and not 100%). Case 2 as defined here includes all the cases with at least two partial response fractions, but includes other cases as well. In the definition of Case 2, one or more animals respond at some dose, such that one or more do not respond at some higher dose (Silvapulle, 1981).

In addition, the geometric mean dose must be higher for animals that respond than for animals that do not respond. The second condition is indicated in Revised UDP as a requirement for inferences regarding the LD50.

In standard probit analysis, bounds of a confidence interval may be infinite. The standard approach for detecting whether the bounds are infinite is based on a test of the statistical significance of the slope parameter. An analogous procedure can be used with the profile likelihood approach.

Case 3. When there are no partial response fractions (along with other requirements of the case, as indicated in Table 2), for technical reasons the profile loglikelihood approach apparently cannot be applied in a straightforward manner. In this case, it seems that any dose within the interval bounded by the highest dose with no responses, and the lowest dose with 100% responses, would be about equally valid as an estimate of the LD50. It seems natural to consider whether those two doses can function in practice as an approximate confidence interval, and there does not appear to be any alternative for defining bounds in this case.

For Case 3, the proposed bounds are not designed to achieve a specific confidence level. Rather, the approach is to ask what is the realized confidence level, if bounds are computed in a certain way.

Case 4. When there is a single partial response (along with other requirements for the case, as indicated in Table 2), the profile loglikelihood can be applied using special computations developed by the ICCVAM Acute Toxicity Working Group. Some technical details are given in Appendix A..

Case 5. This is an infrequent case, which appears to occur primarily when an LD50 is close to a bound. Table 3 is an example of Case 5, generated in a simulation of the Revised UDP.

Table 3. Example of Case 5.

dose (mg/kg)	number tested	number responding
1.0	6	5
1.5	2	2

In the simulations, test doses are restricted to the range 1-5000 mg/kg. For the result displayed in Table 3, testing was probably stopped when three animals tested in sequence at 1 mg/kg all responded.

It could be concluded that the LD50 is more than likely to be below 1.5 mg/kg. A profile likelihood calculation can be done.

2.4. Confidence Intervals Based on Profile Likelihood

This section provides a non-mathematical overview of profile likelihood computations proposed for use when the data from a given study is assigned to Case 2 or Case 4. The methods are illustrated using hypothetical data sets, which were generated in simulations of the Revised UDP.
Some technical details and formulae are provided in Appendix A. The material in this section is not needed in order to understand the evaluation of performance of the methods using simulations, which is found in the sections that follow. However, it is desirable to understand the following points: First, the type of bounds proposed will be infinite in some cases. More precisely, both the upper bound and lower bound will be finite or both bounds will be infinite. This is as in standard probit analysis. Second, the methods proposed cannot be implemented by plugging data into a formula. Specialized computing skills such as numerical optimization are required for implementation. For the numerical aspects, there are multiple alternative algorithms that may be used without actually changing the statistical approach.

Explicit descriptions of the profile likelihood approach are found in Barndorff-Nielsen (1991), Davidson and MacKinnon (1993), and Meeker and Escobar (1995), among other sources. Implicit justification for the approach is found in any theoretical statistics book if it is noted that (I) confidence intervals can be constructed by inverting statistical tests (Section 2.2) and (ii) the method proposed inverts a likelihood ratio test that is ordinarily presented. (These references are somewhat technical. The point here is to confirm that the general type of approach suggested is well established in statistics.) The method has been widely used in connection with nonlinear statistical models, and descriptions can be found in literature associated with various applications. Barndorff-Nielsen (1991) uses the term *profile likelihood* to denote the particular variant of a likelihood function that is used here, while other authors do not specifically name that variant. Barndorff-Nielsen (1991) also reviews refinements of the approach.

According to the approach proposed, statistical results are based on likelihood curves. Figures 3 and 4 provide two examples of likelihood curves, based on hypothetical data examples. Formulae for the likelihood curves are provided in Appendix A. Points to be emphasized

Figure 3. Likelihood Curve for Example 1.

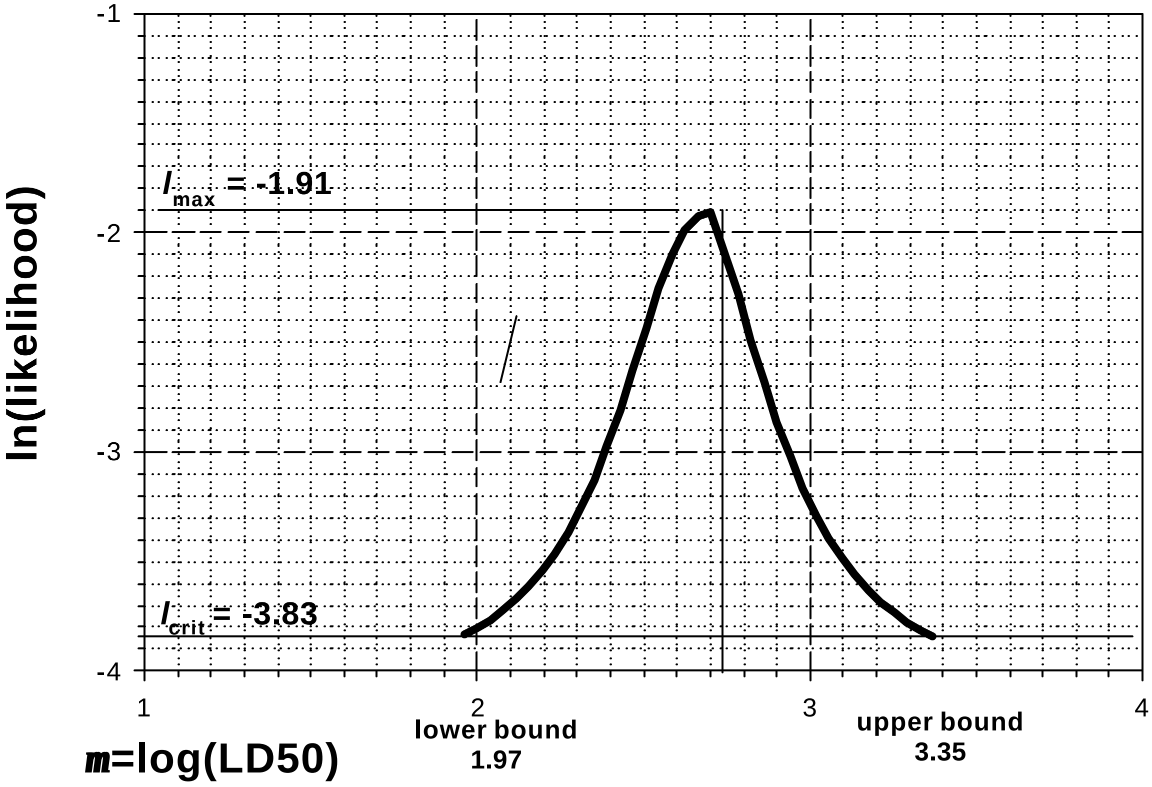
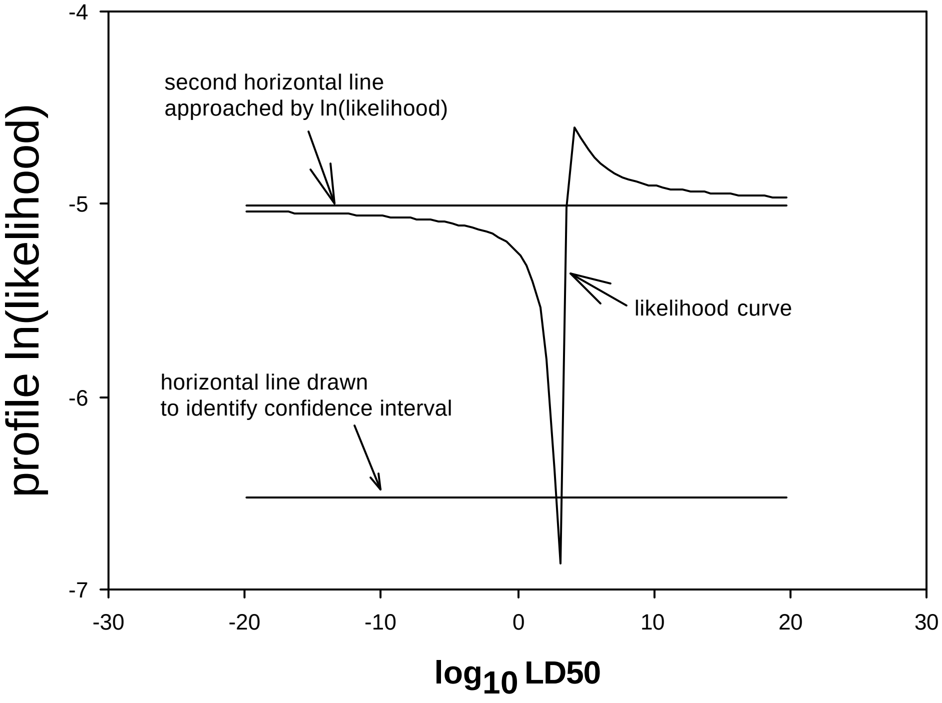


Figure 4. Likelihood curve for Example 2.



include that each distinct data set is associated with a distinct likelihood curve, from which can be read the statistical results (confidence bounds as well as a point estimate) for that data set. The likelihood curve also depends on the type of dose-response function that has been assumed. Revised UDP specifies the use of probit models. A logit model would also have an LD50 and a slope, closely analogous to the probit LD50 and probit slope, but for a given data set the likelihood curves for a probit model and a logit model would not be identical.

For present purposes, it is helpful to think of the likelihood curve as providing levels of relative support that a specific data set give to different choices of parameter values (Edwards, 1992). In particular, the LD50 value with highest likelihood is the *maximum likelihood estimate* (MLE) and may be considered for a point estimate. (However, Revised UDP specifies that the LD50 point estimate will be a MLE based on an assumed slope.) It turns out that standard probit calculations generate maximum likelihood estimates although the likelihood is not computed explicitly (Finney, 1971).

If this notion of likelihood-as-support to calculation of confidence bounds is extended, it seems that values inside the confidence interval should have higher likelihood than values outside the interval. The upper and lower bounds for the confidence interval, it seems, should have equal likelihood (see Figure 3). This notion is the basis of the graphical approach described with the following examples.

Example 1. The following data were generated in a simulation of Revised UDP..

Table 4. Example with a Single Partial Response Dose.

dose (mg/kg)	number tested	number responding
175.0	2	0
553.4	3	2
1750.0	1	1

likelihood values (see text for explanation)
maximized loglikelihood = -1.910
loglikelihood for bounds =-3.830= -1.91 - 1.92
95% CI for the LD50 = 93 - 2258 mg/kg based on method of profile likelihood

Here there is only a single partial-response dose and so standard probit programs cannot be used to generate an estimate of the LD50. The likelihood curve associated with these data is displayed in Figure 3 [the natural log of the likelihood is graphed. Use of ln(likelihood) is conventional in statistics for computations with likelihoods.]

The confidence bounds can be computed graphically using Figure 3, by the following steps:

(1) There are two parameters in the probit model, namely the slope and LD50, but the curve displayed is a function of the LD50 only. A 2-parameter likelihood can be defined which can be graphed in three

dimensions. In the context of Revised UDP, the LD50 is of primary interest. In this context, the slope is said to be a nuisance parameter. Therefore, it does seem useful to obtain a likelihood curve for the LD50 alone, if that is possible.

One way to eliminate the slope, as used in Revised UDP point estimation and stopping rules, would have been to assume a value for the slope. Here, a more computationally intensive approach has been used. The approach proposed is the detail that defines the profile likelihood approach specifically, as a type of likelihood approach. According to the profile likelihood approach, at each value of the LD50 the slope value is used that maximizes the 2-parameter likelihood.

Since the profile likelihood curve is the only likelihood curve that will be used in this document, the profile likelihood (for the LD50 eliminating the slope) will be referred to as "the likelihood curve" although, to be more exact, it should be referred to as the "profile likelihood curve."

(2) For the hypothetical data, the likelihood function has a peak where the log(LD50) has the value of approximately 2.7 (i.e., at an LD50 value of 553 mg/kg). Note that the value of 553.4 mg/kg is the middle dose in this example, the dose with an intermediate response fraction. This value would not be a bad choice of a point estimate for the LD50 for these data.

(3) The subsequent computations require the peak value of the ln(likelihood). In this particular example, special computations are needed to get the maximized (peak) ln(likelihood), which are presented in Appendix A. For the data considered here, these computations yield a value of -1.91 for the maximized ln(likelihood), which is evidently consistent with the curve in Figure 3. In cases where standard probit calculations can be applied, computation of the maximized ln(likelihood) involves a different procedure, as in Example 2 below.

(4) An approximate lower bound for the LD50 can be read from the likelihood curve as follows. A horizontal line is drawn at a (log) likelihood value of -3.83, a value which is computed with a formula below. Referring to Figure 3, this line is seen to intersect the likelihood curve to the left of the curve peak, at an LD50 value of 92 mg/kg ($\log_{10}$(92)=1.965). Therefore, the value of 92 mg/kg is taken to be the lower bound for the LD50.

A similar approach is used for determining the upper bound of 2258 mg/kg ($\log_{10}$ = 3.35). The upper bound value is the dose value where the horizontal line crosses the likelihood at a second point, to the right of the point estimate.

The Y-axis value of the horizontal line (-3.83 for this example) is calculated with the following formula which has been developed by mathematical statisticians:

$$\text{ln(likelihood) for bound} = \text{maximized ln(likelihood)} - 1.92$$

For the example, a maximized ln(likelihood) value of -1.91 has been calculated, so the Y-axis value for the horizontal line is -1.91 - 1.92 = -3.83.

In the formula above, the value of 1.92 is appropriate for computation of a 2-sided 95% interval. A different value would be used to compute a 90% interval, and so on. Technically, the value to be used is taken from tables (or the electronic equivalent) of a chi-square distribution with one degree of freedom.

To see why these computations make sense, reflect again on the notion that the likelihood is a measure of relative support that the data give to alternative choices of an LD50. The graphical approach separates the possible choices of an LD50 into two sets based on their likelihoods: The confidence interval comprises LD50 candidates with ln(likelihood) above the horizontal line, while LD50 candidates outside the confidence interval have ln(likelihood) below the horizontal line. The two bounds are dose values with equal likelihood. The procedure seems natural if LD50 candidates with higher likelihood are regarded as better supported by the data.

(5) Reflection on the procedure just described indicates a possible problem. The likelihood curve was graphed over a finite range. The graphical approach assumes that the ln(likelihood) remains below the horizontal line for LD50 values not graphed. If not, then the bounds are infinite. However, as mentioned previously, there is a way to determine if the bounds are finite or infinite. Use of the formula in this case indicates that the bounds are finite.

Example 2. The following hypothetical data were also generated in a simulation of Revised UDP.

Table 5. Data for Example with Infinite Bounds

dose (mg/kg)	number tested	number responding
175.00	1	0
553.40	2	0
1750.0	3	1
5000.0	4	1

Probit results: slope = 1.02, estimated LD50=14223 mg/kg

Standard probit calculations (Finney, 1971) can be performed in this case. Probit results for the LD50 and slope are displayed in a table footnote. According to standard probit calculations, the bounds for the LD50 are infinite in this case.

The likelihood curve based on these data is displayed in Figure 4.

The curve can be used for the point estimation because the likelihood curve has an unambiguous peak. If the graph is plotted over a more narrow range than that used for Figure 4, it can be seen that the peak actually does correspond to the probit LD50 estimate. In fact, standard probit calculations do generate the maximum likelihood estimates (Finney, 1971).

Next we need the maximized ln(likelihood). In this case, the computations are different from those used for Example 1. When the standard probit calculations apply (as in this example but not in Example 1), the maximized ln(likelihood) is computed by plugging the probit estimates of the slope and LD50 into the two-parameter likelihood formula. The two-parameter likelihood formula is given in Appendix A.

As in Example 1, a horizontal line can be drawn separated from the peak ln(likelihood) by a value of 1.92 units in the direction of the Y axis. The result of this step is the lower of two horizontal lines drawn on the graph (see Figure 4). In this case, although the likelihood curve dips below the horizontal line, the set of dose values with ln(likelihood) above the line (those values not excluded based on our data) stretches to infinity in each direction. Consistent with the results of standard probit computations for this case, the profile likelihood confidence bounds are considered infinite.

Note that if the likelihood curve had been viewed over a narrow range of LD50 values around the peak, one might have concluded that the upper bound was probably infinite, but might be misled to suppose that the lower bound is finite. This problem can be resolved as follows. Observe that in this case as the LD50 approaches infinity in either direction, the likelihood curve approaches a second horizontal line (refer to Figure 4). In fact, it appears that in all cases the likelihood curve will approach some line in this way, and the location of that line can be determined. (The formula is provided in Appendix A.) Evidently, the bounds are finite if and only if the second line is located below the first in the Y-axis direction.

Computer algorithms, particularly handling of infinite bounds. Despite what these examples may suggest, it is not proposed that in practice the bounds will be obtained by literally drawing lines on graphs. A computer program will be used to perform analogous computations. However, understanding of the graphical approach just given can provide an appreciation of the types of computer algorithms required to implement the approach. Three types of specialized computer algorithms are evidently needed.

(I) The approach requires that we compute the maximized value of the ln(likelihood). When the results of a study fall in Case 2, an optimization (peak finding) algorithm is required. Standard probit calculations (Finney, 1971, Ch. 2) represent an appropriate optimization algorithm in this case, and that approach has been used in simulations reported in the following sections.

(ii) Computation of the bounds requires us to identify values of the LD50 that have specific values of the ln(likelihood). For the simulations reported in this document, a bisection algorithm has been used.

(iii) In Example 1, it was explained how the slope is eliminated from the likelihood function when using the profile likelihood method. (For a given value of the LD50, use the slope value that maximizes the likelihood.) Consequently, another optimization routine is needed. In simulations, a type of weighted Gauss-Newton algorithm, also termed a scoring algorithm, has been used. This is a type of optimization method widely used in situations such as probit fitting (Nelder and Wedderburn, 1989).

Each of these three operations involves a kind of iterative search procedure, meaning that some kind of initial guess is developed for a quantity to be computed and that guess is refined in an iterative fashion, until further refinements seem to have no practical effect. The implementation of these types of algorithms requires a specialized type of computing skill.

For each of the three operations identified there are various algorithms that may work. The choice of an algorithm is not fundamental to the statistical method, but can affect the performance of a computer program in some ways. If a relatively poor algorithm does not produce incorrect results, computing speed may be slowed, or the algorithm may occasionally fail to produce results because of a variety of numerical phenomena.

2.5. Simulation Procedures for Measuring the Performance of Confidence Intervals.

In previous work, we have used simulations to evaluate the performance of OECD TG 425 for the purpose of estimating the LD50. In these simulations, values were assumed for the slope, LD50, and starting dose, and numerous data sets were simulated. In that situation, estimates of the LD50 close to the true value are considered desirable. Therefore, performance could be evaluated by considering the percent of simulated studies yielding LD50 estimates in some sense close to the true value, say within some factor of the true value.

Analogous simulation procedures have been used here to evaluate the performance of the proposed confidence intervals. As with previous simulations, values are assumed for the LD50, the slope, and the initial starting dose. For a given combination of assumed values of these parameters, we simulate a large number of studies. The simulation results are used to compute measures of performance. While the procedure for simulating data sets is identical to the procedure used in evaluation of point estimates, different performance indices are computed from the simulated data.

To assess the performance of the confidence intervals, we report four measures of performance, which are denoted PM1, PM2, PM3, and PM4 in the tables of simulation results.

PM1. This is the estimated percent of studies that have finite confidence bounds. (The bounds are both finite or both infinite.) It is desirable to have narrow confidence bounds, but it is not clear that the occurrence of very wide bounds should be viewed as a drawback for the method of computing confidence bounds, versus as a drawback of the study design. In any case, the index seems to provide useful information.

PM2. This is the coverage, which is the fraction of studies for which the true LD50 falls inside the confidence interval (above the lower bound and below the upper bound). For each of, say, 1000 simulated studies, the confidence intervals are computed with the procedures proposed, and the study is scored as either enclosing the true LD50 or otherwise. PM2 is then the percent of the 1000 simulated studies with bounds that enclose the true LD50. In cases where the bounds were infinite, they were scored as enclosing the true LD50.

By the definition of a 95% confidence interval, the ideal value for PM2 is 95%. Ideally, PM2 will not vary when the LD50, slope, and initial test dose are varied.

PM3. PM3 and PM4 are alternative measures of the typical widths of confidence intervals. PM3 is the median ratio of the upper bound to lower bound. (The ratio upper/lower is computed separately for each of, say, 1000 simulated studies. PM3 is then the median of the 1000 ratios.)

In cases where the bounds were infinite, the ratio was coded as greater than 1000 (>1000). Ratios that were finite but greater than 1000 were also coded as simply >1000. (Otherwise some confidence intervals with finite bounds would be coded as more narrow than intervals with infinite bounds.) For many of the situations where a slope of 0.5 or 0.8 was simulated, over 50% of simulated studies had infinite slopes (that is, PM1>50%). (See Table B.2 of Appendix B.) In these cases , the value of PM3 is >1000. In a few cases, PM3 was >1000 when PM1 is slightly below 50%, because of some intervals that are very wide but not infinite.

Use of a value of 1000 is somewhat arbitrary but this choice does not effect the median ratio unless the ratio exceeds 1000 for at least 50% of simulated studies. We suggest that if the median ratio is greater than 1000, there is not much practical value in quantifying the proportion of confidence intervals with infinite bounds, versus with bounds that are finite but separated by a factor of 1000 or more.

In several cases where a steep slope is assumed, PM3 is equal to 3.2. This is the ratio of adjacent test doses, except in some cases where a test dose is a limit dose. In these cases, many data sets fall under Case 3, for which all doses are associated with either 0% response or 100% responses..

PM4. This is a second measure of typical width, the median standardized width of the confidence interval. For each simulated study (say for 1000 simulated studies), we compute the quantity:

$$\text{standardized width of confidence interval} = 100 * \frac{\text{upper bound - lower bound}}{\text{true LD50}}$$

This ratio is computed for each of, say, 1000 simulated studies. PM4 is then the median of the 1000 standardized widths.

In cases where the bounds were infinite, the standardized width was scored as >100,000. This is comparable to use of a code of >1000 for PM3 given the approximate relationship between the two indices.

In interpreting these measures, it may be useful to consider the coverage measure PM2 jointly with measures of width (PM3 or PM4). If the coverage is larger than 95% and the intervals appear undesirably wide, then a case can be made for refining the statistical procedure to yield more narrow bounds, with coverage closer to the ideal value of 95%.

OECD standard simulation scenarios for acute mammalian guidelines. Simulations have been conducted based on two sets of scenarios. (the term scenario is used to mean a combination of true LD50, true slope, and initial test dose.)

The first set of scenarios comprised 45 combinations of slope and LD50, with the initial test dose set to 175 mg/kg in each case. The value of 175 mg/kg is the Revised UDP default initial test dose, to be used when there is no reliable information to indicate a better initial test dose. The combinations of slope and LD50 for this set are the same as for the second set.

The second set of scenarios comprises 112 combinations of slope, LD50, and initial test doses. This set of scenarios has been developed by OECD workgroups for evaluation and comparison of acute toxicity designs. For this set, initial test doses were initially specified in terms of percentiles of the tolerance distribution. The test doses were then calculated from the slopes and LD50s. In simulations of the Revised UDP, test doses are restricted to the range of 1 to 5000 mg/kg. Therefore, combinations with an initial test dose outside that range have been deleted.

In this set, scenario number 95 has been modified for the simulations reported here, by changing the initial test dose from 4870 to 4750 mg/kg. The LD50 is 3000 mg/kg for this scenario so that testing tended to be concentrated on the two doses 4870 and 5000 mg/kg, the latter being the limit dose. The original value of 4870 mg/kg is unrealistically close to the limit dose of 5000 mg/kg and the scenario was unmanageable numerically because of a large number of numerical overflows. When the initial test dose was changed to 4750 mg/kg, no further difficulties were encountered. (No numerical problems were encountered with any of the other scenarios, after some refinements of the algorithms.)

Additional details of simulation. The performance measures PM1-PM4 were computed only using data for Cases 2-4, because it is only for those cases that statistical methods are proposed in this document. For example, PM1 is then the percent of studies in Cases 2-4 that have finite intervals. However, the percentages of studies assigned to different cases were computed using the data for all cases.

For each scenario, a minimum of 1000 studies was simulated. Because confidence intervals were computed for Cases 2, 3, and 4, the combined number of simulated studies for those 3 cases was fixed at 1000 for each scenario while the total number simulated studies per scenario was variable but always greater than 1000.

As in previous simulations, the range of test doses has been restricted to the range of 1 to 5000 mg/kg.

2.6 Simulation Results

Two types of simulation results are provided in Appendix B.

Table B.1 of Appendix B provides percentages of Cases 1-5 for each scenario. (See Table 2 of this

Section for the definitions of these cases.) A combined percentage is reported for Cases 1 and 5. Table B.1 contains the results for both sets of scenarios, those with the initial test dose fixed at 175 mg/kg and those with initial test dose varied.

The case frequencies are informative regarding how often particular procedures can be applied. In particular, the low frequency of Case 2 in many scenarios supports our assertion that standard procedures of probit analysis will often not be applicable with TG 425. Cases 1 and 5 occur with relatively high frequency when the true LD50 is close to a limit dose. This is probably a consequence of instances where a particular stopping rule is invoked, namely that testing is stopped if three animals tested in sequence at 5000 mg/kg do not respond, or if three tested in sequence at 1 mg/kg all respond.

The relative frequencies of different cases depends strongly on the slope, for obvious reasons. If the slope is steep, then the percentages of animals responding changes from 0% to 100% within a narrow range of dose values, and the possibility for obtaining a partial response percentage therefore relatively small.

Table B.2 provides the values of performance measures PM1-PM4 (defined in Section 2.5) for each Scenario. Overall, the results seem to suggest acceptable performance of the methods proposed.

The results indicate a strong dependence on the slope. As the slope increase, the percentage of infinite bounds is lower (PM1), the coverage increase (PM2), and the intervals become more narrow (PM3, PM4).

With regard to coverage (PM2) the ideal value is 95%, and ideally the coverage will not depend on the slope. Therefore, the PM2 values of 99%-100%, associated with steep slopes, are not necessarily to be viewed favorably. However, in the steep-slope situations, the confidence intervals tend to be narrow (PM3, PM4). Thus, the conservatism of the methods when the slope is steep (as quantified by PM2) do not seem to represent a serious drawback of the methods proposed.

3.0 Software

3.1 Purpose and Description

Because the Revised UDP is relatively complex statistically, dedicated software has been developed to integrate all statistical features of the test, including a) multiple stopping criteria; b) estimation of an LD50; and c) provision of confidence intervals , together with their appropriate places in the laboratory protocol. This software was developed for a Windows environment and is accompanied by a user manual. The software and manual are designed to be readily understood and implemented by scientists outside the U.S. who may have limited facilities and English comprehension. It will be a stand-alone package designed for analysis only, with provision for an investigator to create reports that include animal identifiers that match those in a laboratory's standard data maintenance files, thereby permitting data verification.

Development of this software is being carried out under contract to the U.S. EPA, through work assignments 4-06 and 5-03 of Contract No. 68-W7-00285. Building the package follows practice for verification, which is an abbreviated form of standard practice such as that outlined by the FDA draft guidance for industry on general principles for software validation. The FDA guidance states:

> “Verification is defined in 21 CFR 820.3(aa) as "confirmation by examination and provision of objective evidence that specified requirements have been fulfilled." In a software development environment, software verification is confirmation that the output of a particular phase of development meets all of the input requirements for that phase. Software testing is one of several verification activities, intended to confirm that software development output meets its input requirements. Other verification activities include walkthroughs, various static and dynamic analyses, code and document inspections, both informal and formal (design) reviews and other techniques”.

The model of verification is not unlike the QA/QC Check of the Benchmark Dose System (BMDS) Software for the U.S. EPA (Contract No. 68-C9-8007, Work Assignment 1-10, December 1999).

Completion of all construction, testing, and documentation is scheduled for summer 2001.

3.2 Quality Assurance/Quality Control

Software requirements are being set out by the U.S. EPA and the contractor regarding environment, input/output/functions, user interfaces, error handling; design is considering implementation (coding) issues; and testing will be performed to ascertain that the package does what it is designed to do. Some of this testing will be in the form of stressing the program by pushing it to unusual circumstances (and sample data sets are currently under construction). Some of these data sets generally can be described by the case descriptions in section 2 of this document. The sets specifically encompass, however, such situations as possible data entry errors and the various stopping circumstances, as well as unusual dose magnitudes. Some of it will constitute simulations characterizing the behavior of Revised UDP that can be compared to independently programmed output regarding Revised UDP behavior. When completed, these activities will constitute a verification of the analysis package.

At the first stage, an outline of the program has been created, identifying its structure (with data, calculation, and report modules, and, for testing, a simulation module), how modules will interact, what each module will do and, as appropriate, the mathematics for those operations; enumerating the possible configurations of data and which will and will not give numeric solutions; describing messages (prompts, warning, error) from package to user and their circumstances; and outlining the testing and simulation processes. Concurrently, an outline of the user manual was delivered.

4.0 References

Barndorff-Nielsen, O.E. 1991. Likelihood theory. Chapter 10 in D.V. Hinkley, N. Reid, and E.J. Snell (eds) *Statistical Theory and Modelling*. Chapman and Hall.

Bickel,P.J. and K. A. Doksum. 2001. *Mathematical Statistics: basic ideas and selected topics*. Volume 1. (2nd ed.) Prentice Hall.

Casella, G., and R.L. Berger. 1990. *Statistical Inference*. Duxbury.

Edwards, A.W.F. 1992. *Likelihood*. 2nd ed. Johns Hopkins Univ. Press.

Davidson R., and MacKinnon, J.G. *Estimation and Inference in Econometrics*. Oxford U. Press.

Finney, D.J. 1971. *Probit Analysis*. 3rd ed. Cambridge U. Press.

McCullagh, P., and J.A. Nelder. 1989. *Generalized linear models*. (Second ed.) Chapman & Hall/CRC.

Meeker, W.Q., and Escobar, L.A. 1995. Teaching about confidence regions based on maximum likelihood estimation. The Amer. Statistician. 49(1):48-52.

Silvapulle, M.J. 1981. On the existence of a maximum likelihood estimators for the binomial response model. J. Royal Statist. Soc. Series B 43(3):310-313.

Appendix A

Performance Characteristics of the Revised UDP Point Estimate and Confidence Interval

1.1 LD50 Confidence Bounds for Revised UDP: Technical Specifications and Numerical Programming

This appendix provides technical detail and mathematical formulas, and supports technical peer review and programming.

The preliminary approach, described in this Appendix, was to limit the numerical search for a bound to a finite interval above or below a point estimate of the LD50. This approach was used because no procedure was readily available to determine from the data, *a priori*, whether the bound is finite or infinite. It was suggested that the search interval may be made sufficiently wide so that, if a bound is outside the interval, it might be considered infinite for practical purposes.

However, it appears that there is actually a criterion that can be used to determine whether the bounds are finite or infinite. The probit model can be parameterized in terms of $\mu = \log_{10}(\text{LD50})$ and the slope (\$). According to the method of profile likelihood, the decision of whether a value of μ is inside or outside the confidence interval is made by optimizing the slope parameter with μ fixed at the value of interest, and thus obtaining an optimized loglikelihood value corresponding to a particular μ value. The value of μ in question falls within the confidence region if and only if the maximized loglikelihood is greater than or equal to a critical loglikelihood that can be denoted as l_{crit}. The computation of l_{crit} is as described in this Appendix.

As the value of μ is taken toward infinity in either direction, and the slope is optimized for each value of μ, the optimized slope value is observed to converge to zero. The loglikelihood is observed to converge to a value that can be computed directly, by substituting for each predicted response percentage the pooled response percentage $p_{\text{pooled}} = \sum_{i=1}^{g} r_i \Big/ \sum_{i=1}^{g} n_i$ where r_{i} and n_{i} are the numbers of animals that respond and the number tested at the ith of g dose levels tested.

This behavior can be understood as follows. For definiteness, consider computation of the lower bound. As μ is taken toward negative infinity, the value of \$ approaches zero. If \$ did not approach zero, then all of the predicted response probabilities would go to zero. However, the methods are applied only when some animals respond and others do not. Therefore, \$ goes to zero to fit a mixture of animals that responded and did not respond as μ is taken to infinity. When \$ is close to zero, the doses become, for purposes of probit analysis, about the same dose. (For purposes of probit analysis, the magnitude of dose ratios is considered relative to the slope.) Therefore, as μ is taken to infinity and \$ optimized at each value of μ, the fitted probit line approaches a line connecting the point (μ,probit(0.5)) to the point ($\bar{x}$,probit(p_{pooled})) where $\bar{x}$ is the mean log dose.

Therefore, the criterion for determining whether or not the bounds are finite is as follows. Let l_{pool} denote the value of the loglikelihood computed with each response percentage set equal to p_{pooled}. Since the profile loglikelihood will approach l_{pool} as μ approaches ± 4, the bounds are finite if and only if l_{pool} is less than l_{crit}.

Figure 3 in Section 2.4 of this document, and the associated discussion of Example 1 is misleading. If the loglikelihood is graphed over a sufficiently wide range of doses, the loglikelihood is seen not to be convex and the nonlinear equations that define the bounds have more than two roots. (In the graphs of Section 2.4 of this document, the curve crosses the line more than twice.) In this case, according to the criterion just described, the lower bound as well as the upper bound is infinite, which is also the result obtained with the standard probit methods.

1.2 Background. The ATWG proposes to implement confidence bounds for the LD50, for use with acute toxicity data generated in accordance with the Revised UDP. The method for calculating the confidence interval will be available in software developed to support the Revised UDP; this software will also provide point estimates of the LD50 and will evaluate stopping criteria. The decision to develop new confidence interval procedures is based on simulations that indicate that standard procedures (for analysis of data under a 2-parameter probit model) will very often not be applicable with data generated according to the Revised UDP. This Appendix is intended to support statistical peer review of confidence interval procedures, and (subject to modifications based on the review) to support numerical programming.

Based on simulations presented in Section 2 of this document, it appears that in most cases it will be possible to compute a confidence interval using one of two procedures and that these procedures will have acceptable performance.

In cases where no animals respond at some doses, and all animals respond at some other doses (the latter being greater than the former), it is proposed that the lower bound for the LD50 will be the highest dose associated with no observed response. Similarly, the upper bound will be the lowest dose associated with responses for all animals tested at the dose.

In most other cases, it will be possible to compute a bound using the method of profile likelihood (see Barndorff-Nielsen, 1991, Section 10.2.4). In particular, it is proposed that this approach will be used in most cases where there is only one dose with an intermediate response fraction (neither 0% nor 100% responding), a case that is not handled by standard probit methods. (Proposals for handling various cases are summarized in Section 1.5 of this Appendix.)

Of the two procedures, the profile likelihood approach is the primary focus of this Appendix. The approach requires handling of a number of special cases and specification of other technical details.

Although a description of the profile likelihood approach has been included here, this document is intended to be reviewed primarily by individuals with some background in likelihood based statistical procedures. In addition, it is assumed that readers are familiar with certain types of numerical techniques (line searching and optimization) as used in implementation of nonlinear statistical models.

The material which follows is organized into three sections.

Section 1.3 presents notation, the probit dose-response model, and the profile likelihood approach for computation of confidence intervals. Comments are provided on alternative parameterizations of the probit model.

Section 1.4 discusses numerical algorithms. Three types of specialized numerical routines are required: 2-dimensional optimization to calculate maximum-likelihood estimates, line searching to compute bounds, and 1-dimensional optimization (nested within the line search).

Section 1.5 presents a classification of cases, with proposals regarding how each case is to be handled. Different cases require different confidence interval computations and, for some low-frequency cases, confidence intervals are not proposed.

1.3 Overview of parametric approach

Notation for describing grouped data. For present purposes, it is convenient for the data to be summarized by dose level. Let:

g = number of dose levels tested;
d_i = ith dose level evaluated, $I = 1,...,g$. We assume that d_1 is lowest test dose, d_g is the highest, and so on.
x_i = $\log_{10}(d_i)$
n_i = number of animals tested at the ith dose level, $I = 1,...,g$;
r_i = number of animals observed to respond at the ith dose level, $I = 1,...,g$.

While data summarized in this way are convenient for the computations described here, some computations associated with the stopping rules cannot be calculated from data summarized in this way.

Probit dose-response model. A probit curve is fitted to the data, relating the fraction of animals that respond and the logarithm of dose. The probit model has two parameters. According to one parameterization (the parameterization proposed for final results), the probity parameters are the *slope* (say \$) and the LD50. For purposes of this document, it is convenient to make use of the parameter $\mu=\log_{10}($ LD50). For likelihood-based statistical procedures such as those used here, it is permissible to do estimates and confidence intervals directly for μ and then transform those results to results for the LD50.

Let $p(x ;\mu, \$)$ denote the probability of response, where x is the common logarithm of dose. Then an expression for the probity model is:

$$p(x\,;,\,\$) = M[\,(x - \mu)\cdot\$\,]$$

where $M(z)$ denotes the cumulative distribution function (CDF) for a standard normal distribution.

Calling the parameter $\$$ a "slope" is a toxicological convention. Probity analysis is commonly described as a linear regression of a transformed response (probity percentage response) against the logarithm of dose. To see this, rearrange the expression above as follows:

$$M^{-1}[\,p(x\,;\mathrm{LD50},\$)\,] = \$\cdot x - \$\cdot\log_{10}(\mathrm{LD50}\,)\,)$$

where M^{-1} denotes the inverse of function corresponding to M, so that evidently the relationship between dose and response can be transformed to a linear relationship with slope $\$$ and intercept -$\$\log_{10}$(LD50)).

Note the use here of the common (base-10) logarithm of dose, which is a toxicological convention. For some purposes, the choice of a base for logarithms is arbitrary, but the common logarithm needs to be used in software designed to support Revised UDP, in order to have comparability of results obtained with different programs. In particular, the value of the slope estimate will depend on the base chosen for logarithms.

An alternative parameterization, associated with a particular interpretation of the probity model, is:

$$p(x\,;\mu,F) = M[\,(x - \mu)\,/\,F\,]$$

where $\mu = \log_{10}(\mathrm{LD50})$ and $F = 1/\$$. Of course, μ and F^2 are conventional notation for the mean and variance of a normal distribution. This parameterization may be preferred particularly when the probity model is interpreted in terms of a *tolerance distribution*. According to that interpretation, variation among test animals in response to a particular dose is related to individual variation in sensitivity to the test substance. The tolerance of a single individual is defined to be the dose that will cause that individual to respond, given its sensitivity to the test substance. Then the fraction responding at a given dose equals the fraction of individuals with tolerance below that dose. A frequency distribution is assumed for variation of tolerances among individuals. The probity formulae result from assuming a lognormal distribution for tolerances, with parameters μ and F.

For purposes of the procedures described in this Appendix, the $\mu,\$$ parameterization has proved to be more convenient than the μ, F parameterization. In particular, it appears that widely different values of F can be associated with slope values about equal to zero, and log-likelihood values that are not much different.

Point estimation of the LD50. This Appendix is concerned primarily with interval estimates rather than with point estimates. However, the following remarks may help to place in perspective the various computations that need to be implemented in the software. The purpose of acute testing under the Revised UDP is to obtain an LD50 estimate. In this context, the probity slope is a nuisance parameter. Revised UDP specifies that when estimating the LD50, a value will be assumed for the slope parameter (the default assumption is a slope of 2) and that the LD50 will be estimated based on the resulting 1-parameter model using maximum likelihood. Revised UDP provides an expression for the likelihood function. The LD50 point estimate is not used in the computations for the confidence interval developed in this Appendix. Computations for the Revised UDP stopping rule also involve a distinct point estimate of the LD50, for different reasons.

Two-parameter and profile log-likelihood functions for grouped data. Likelihood functions are functions of model parameters, which are used in statistical inferences about those parameters. Each distinct data set yields a distinct likelihood function. It can be helpful to think of a likelihood function as measuring the relative support that the data provide for alternative choices of parameter values, with higher loglikelihood values indicating relatively stronger support. For example, the maximum-likelihood estimates of the parameters µ and \$ are the parameter values that maximize the 2-parameter function. The exact roles of these functions in computation of confidence intervals are described in detail below.

The following two likelihood functions need to be defined for the methods proposed. The log-likelihood function for the two-parameter probity model is:

$$l(m,b) = \sum_{i=1}^{g} \{r_i \bullet \ln(p(x_i;m,b)) + (n_i - r_i) \bullet \ln(1 - p(x_i;m,b))\}$$

(Note the use here of the natural logarithm rather than the common logarithm, which contrasts with the transformation of doses.)

Here, statistical inferences will focus on µ, whereas \$ will be treated as a nuisance parameter. In this context it is useful to define a type of loglikelihood that is a function of µ only, with \$ eliminated. The *profile loglikelihood* function is:

$$l_P(m) = \sup_b l(m,b)$$

In words, define the profile loglikelihood function to be the function of µ only, obtained by setting \$ equal to that value which maximizes the 2-parameter likelihood *l*(µ,\$), fixing µ. This requires a numerical optimization (numerical techniques are described in the next section). In practice the profile likelihood is handled using the same procedures as the likelihood of a single-parameter model, *e.g.*, in likelihood ratio tests (Barndorff-Nielsen, 1991).

Confidence intervals based on profile log-likelihood, "basic" approach. For the likelihood-based intervals considered here, the interval is the set of parameter values not rejected using a likelihood ratio test. The procedure can be stated most simply in the case where unique, finite maximum likelihood estimates (MLEs) exist for both probity parameters, in the interior of the space of allowable values. In this case the approach is fairly straightforward.

Let ^ be the MLE for μ and let $\hat{b}$ denote the ML for \$, which is to say that $\hat{m}$ and $\hat{b}$ are the choices of parameter values that maximize the likelihood function. Then the *maximized value of the log-likelihood* , say l_{sup}, is obtained by plugging the MLEs into the likelihood expression. Thus:

$$l_{\text{sup}} = l(\hat{m},\hat{b}) = l_P(\hat{m})$$

(Here "sup" is short for "supremum.") Then, for a 2-sided 95% confidence interval, the upper bound and lower bounds for μ, say $\overline{m}$ and $\underline{m}$, are obtained by solution of the following nonlinear equations:

$$l_P(\underline{m}) = l_P(\overline{m}) = l_{\text{sup}} - 1.921, \quad \underline{m} < \hat{m} < \overline{m}\ .$$

In general, to compute a 100(1 - ")% confidence interval, the bounds are defined by the equation:

$$l_P(\underline{m}) = l_P(\overline{m}) = l_{\text{sup}} - \frac{1}{2}c_1^2(1-a), \quad \underline{m} < \hat{m} < \overline{m}$$

(Bickel and Doksum, 2001) where $c_1^2(1-a)$ denotes the (1- ")th quantile of a chi-square distribution with a single degree of freedom. (In particular $c_1^2(0.95) = 3.84$=2*1.92.) It is useful to define

$$l_{\text{crit}} = l_{\text{sup}} - \frac{1}{2}c_1^2(1-a)$$

which is the critical value of the profile loglikelihood that the bound values must satisfy.

Use of these expressions requires numerical searches among values of μ above and below $\hat{m}$. In some cases a solution does not exist, in which case the bound may be taken to be ±4. In particular cases, graphs of the profile likelihood suggest an approach to an asymptote that falls short of the critical value. Unless conditions can be derived and automated for identifying the apparent infinite-bound cases, the search must be restricted to a finite interval. When the search is restricted to a finite interval, one cannot distinguish between bounds that are very wide and bounds that are actually infinite.

Example. The following hypothetical data were generated in a simulation of the Revised UDP. The profile loglikelihood curve for these data is displayed in Figure A.1

Table A.1. Data for Profile Loglikelihood Example

dose (mg/kg)	number tested	number responding
175.0	1	0
553.4	2	0
1750.0	3	1
5000.0	4	1

MLEs: $\hat{m}$ = 4.153, $\hat{b}$ = 1.020, estimated LD50=14223 mg/kg

95%CI for LD50 (1950 mg/kg,>2*10^5 mg/kg)

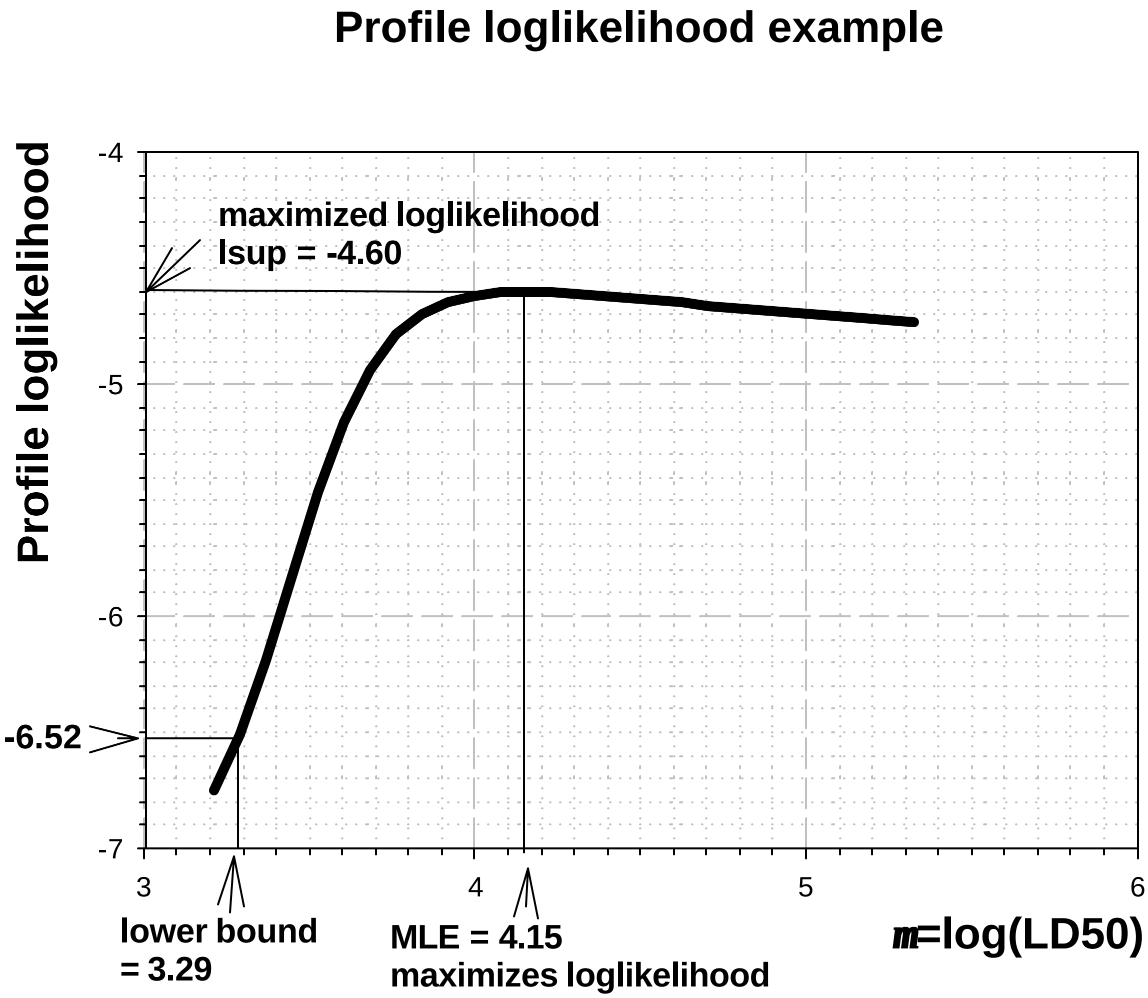
maximized loglikelihood: l_{sup} = -4.603

critical loglikelihood for bounds: l_{crit}=-6.524

This data set was analyzed in the following steps. The 2-parameter probit model was fitted to the data using a conventional probit methods (weighted Gauss-Newton optimization). That approach is considered to yield MLEs of model parameters in this case. The MLEs are displayed in table footnotes.

Evaluation of the loglikelihood at the MLEs gives l_{sup}=-4.603 (see Figure A.1). Therefore, any bounds must have profile loglikelihood equal to l_{crit}.=-6.524. A line search below the MLE found the lower bound for μ of 3.29 (or LD50=1950). A search for an upper bound failed to find a value of μ with the required profile loglikelihood within a factor of 15 of the MLE. Therefore, the upper bound would be reported as greater than 213000 (=15*14223). In this case, the absence of a useful upper bound probably results from the restriction of test doses to values not exceeding 5000 units

Figure A.1 Profile Loglikelihood Example



Extension of the approach to cases with a single intermediate response fraction. In some cases, the computations just described will not be applicable. However, the approach has been extended to one case that is not ordinarily analyzed under a 2-parameter probit model. This is the case where there is only a single test dose with an intermediate response fraction (the percentage responding is neither 0% nor 100%), and where any lower test doses are associated with 0% response, and any higher test doses are associated with 100% response. This is Case 4 as described in Table 1.

In the cases considered above, the loglikelihood supremum l_{sup} was found by evaluating the loglikelihood at the MLEs. For Case 4, it appears that l_{sup} has a natural definition, although the value of l_{sup} is obtained as a limit and does not correspond to particular finite values of μ and $\$$.

Within the ranges allowed for μ and $\$$, the fitted probit curve can be made to match the data as closely as we like by specifying $\$$ to be sufficiently steep. Consider the family of curves that exactly match the single partial response, with different slopes. Steeper slopes allow the 0's to be matched more closely at one end, and the 100's to be matched more closely at the other end. This argument suggests that the supremum of the loglikelihood can be calculated by taking appropriate limits, resulting in the expression:

$$l_{\text{sup}} = r_j \bullet \ln\left[\frac{r_j}{n_j}\right] + (n_j - r_j) \bullet \ln\left[\frac{n_j - r_j}{n_j}\right]$$

where j is the index of the dose associated with a partial response fraction. This expression is obtained from the 2-parameter log-likelihood $l(\mu,\$)$ by deleting the contributions from doses other than dose j, and for dose j by setting the predicting response percentage equal to the observed response percentage $r_{\text{j}}\,/n_j$. For the terms other than the jth, the limit is zero as the slope is taken to infinity. The jth observed response fraction equals the corresponding observed fraction because for any finite slope value, the intercept can be adjusted so that the results for the jth dose are matched exactly.

A second requirement for implementation of a profile likelihood approach is to define a finite interval of μ values in which to search for an upper or lower bound. Where there is an unambiguous MLE, an upper bound is searched for among values of μ above the MLE, and a lower bound is searched for below the MLE. In the case under consideration, where there is a single dose with partial response, we use the dose that has partial response as a bound for the search interval.

Example with one partial response dose. The following data were generated in a simulation of Revised UDP.

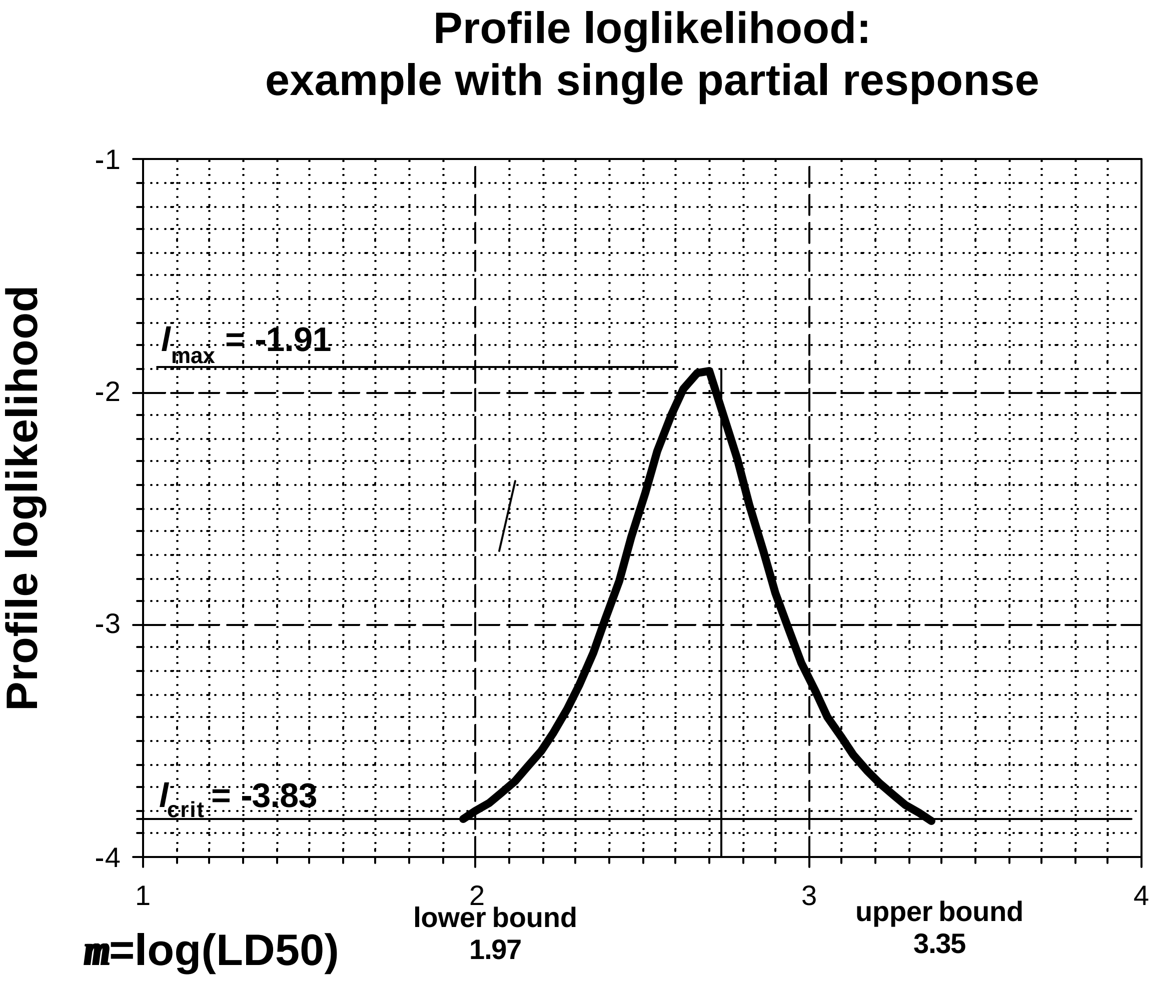
Table A.2. Example with a Single Partial Response Dose.

dose (mg/kg)	number tested	number responding
175.0	2	0
553.4	3	2
1750.0	1	1

maximized loglikelihood: l_{sup} = -1.910
critical loglikelihood for bounds: l_{crit}=-3.830
95%CI 1.97 - 3.35 for μ, 93 - 2258 mg/kg for the LD50

In this example, there is only a single partial response dose. The maximized loglikelihood is calculated using the formula given. (The LD50 would be 553.4.) The graph of the profile loglikelihood (Figure A.2) does not suggest any problem with this way of defining l_{sup}. Each point plotted corresponds to specific finite values of the parameters, but nevertheless the proposed method for calculating l_{sup} (which does not correspond to any particular parameter values) appears consistent with the rest of the curve. The use of such a profile loglikelihood presents no obvious problems.

Figure A.2. Profile Likelihood: Example with Single Partial Response.



1.4 Numerical Algorithms for Likelihood Calculations

A number of technical decisions are required in order to implement a profile likelihood procedure. It is desirable first of all to have criteria for determining if a 2-parameter maximization of the loglikelihood can be performed. In that case, the parameter values that maximize the loglikelihood are the MLEs. In any case, the computation of a bound for the LD50 requires a line search of a finite interval. Some procedure is needed to define the interval that will be searched for a bound. The line search involves evaluation of the profile loglikelihood function $l_P(\mu)$ for different values of μ. Each evaluation of the profile loglikelihood involves a
1-dimensional optimization (\$ is optimized with μ fixed).

Each of these procedures requires a number of technical decisions. Most of these decisions are not related to the fundamental method, being more to production of a reliable algorithm. Here, a description is provided of the implementation used. In simulations, it appears that this algorithm never aborts because of numerical overflows or divisions by zero, etc. For concreteness, the procedure is described for computing the upper bound. The modifications needed for computation of a lower bound seem obvious in most cases.

Computation of MLEs by 2-dimensional optimization. When an optimum can be determined for the likelihood function, the results are used in calculating bounds by the profile likelihood method. There are many optimization techniques that can be considered for this purpose. In probit analysis, it is conventional to use a weighted Gauss-Newton approach devised by R.A. Fisher. This algorithm is described in Finney's (1971) Chapter 4. The approach is considered to generate maximum likelihood estimates in probit analysis. This algorithm is considered to be a perfectly good approach viewed from the standpoint of modern nonlinear statistical modeling. The algorithm is actually a special case of an approach widely used for generalized linear models, a broad class of nonlinear models (McCullagh and Nelder, 1989). The algorithm is closely related to the more familiar Newton-Raphson algorithm, but involves a simplified expression for the Hessian.

It is known that finite MLEs do not exist in some cases. Silvapulle (1981) has presented necessary and sufficient conditions for existence of MLEs for logit and probit models. The conditions are very general, addressing models with many regressors. In the case of probit analysis, the conditions apparently reduce to a requirement that some dose where one or more animals respond is lower than some other dose where one or more animals do not respond. A particular case of Silvalpulle's condition is the case where there are at least two doses with partial response fractions. The latter is sometimes used as a criterion for when probit analysis can be performed. Another case is when the observed relation between dose and fraction responding deviates from monotonicity.

Silvapulle's criterion allows an estimate if the probit slope equal to zero. If the slope is zero, the same response fraction is predicted at every dose. In that case either there is no estimate of the LD50 or else every dose is estimated to be the LD50. A great many applications of probit or logit models are not concerned with estimation of an LD50, and Silvapulle in particular does not discuss estimation of the

LD50.

Currently, 2-dimensional optimization is performed when the Silvapulle condition holds and when an additional criterion is met, which indicates a positive relationship between dose and observed response. (The handling of various cases is summarized in the following section.) In addition to the Silvapulle condition, a requirement is that the geometric mean dose is higher for animals that respond than for animals that do not respond. This condition is indicated in Revised UDP as a requirement for inferences regarding the μ.

<u>Specification of interval searched for a bound</u>. A numerical search for a bound for the LD50 must be restricted to a finite interval, particularly in view of the possibility that a bound may be infinite. The *search interval* is defined using two numbers, a point estimate and a multiplicative factor, say F_{search}. For computation of the upper bound, the search interval is

$$[\text{LD50 point estimate, LD50 point estimate* } F_{\text{search}}].$$

For the lower bound, the search interval is

$$[\text{LD50 point estimate / } F_{\text{search}}, \text{ LD50 point estimate}].$$

With regard to notation, the usual practice of using μ (=$\log_{10}$(LD50)) instead of LD50 is deviated from. This is because, in the software, it is expected that all results will be expressed in terms of the LD50. The variable F_{search} will be accessible for modification for the user. Therefore, F_{search} is represented as a multiplicative factor applied to a point estimate of the LD50.

Here, use of the term “point estimate” is possibly the source of some confusion. The LD50 value which defines the search interval is not the LD50 point estimate indicated in Revised UDP. Therefore, the term “center of search interval” may be used in the remainder of the document. To avoid having to define additional symbols, $\hat{m}$ will continue to denote the center of the search interval although, in statistics, the “hat” (ˆ) over a parameter symbol ordinarily indicates a maximum likelihood estimate.

<u>Determining if a bound exists within the search interval (bracketing step)</u>. The line search algorithm has two steps, a *bracketing step* and a *bisection step*. The bracketing step serves to determine whether a bound exists within the search interval. Also, the bracketing step produces quantities useful in the bisection step, which follows.

Expressing the model in terms of μ, the search interval for an upper bound can be denoted ($\hat{m}$, $\hat{m}$ + $\log_{10}$(F_{search})). A bound exists within the search interval provided that
l_P($\hat{m}$ + $\log_{10}$(F_{search})) < l_{crit}. If this condition holds, then the bisection step can be used to locate the bound value within the search interval. Otherwise, the upper confidence bound is reported only as being greater than the bound of the search interval, i.e., as greater than

$\hat{m}$ + $\log_{10}(F_{\text{search}})$.

This suggest that the bracketing step need only involve evaluation of the profile loglikelihood at the bounds of the search region. However, a more complex set of computations is used: Observe that evaluation of l_P involves optimization of \$. A starting estimate of \$ is required for each optimization. Therefore, it is reasonable to evaluate a sequence of μ values $\hat{m}$, $\hat{m}$ +*, $\hat{m}$ +2*, ..., where * is some constant, stopping when the value of l_P is less than l_{crit} or the bound of the search region is attained. If this approach is used, then good starting values of \$ are usually available. The optimized value of \$ from one evaluation of l_P is a good starting value for use in the next optimization.

In simulations, $F_{\text{search}} = 50$ is used currently, and a value of * is used such that the bound of the search region is attained in 40 steps.

Calculation of a bound by bisection. The use of bisection to calculate a bound for μ requires two values, say μ_1 and μ_2, that satisfy $l_P(\mu_1) > l_{\text{sup}}$ and $l_P(\mu_2) < l_{\text{sup}}$. Such values are provided by the final two values of μ evaluated in the bracketing step.

Gauss-Newton algorithm to optimize \$ with μ fixed. The profile loglikelihood function $l_P(\mu)$ is a function of μ obtained from the 2-parameter loglikelihood $l(\mu, \$)$ by optimizing \$, with μ fixed.

The Gauss-Newton approach, conventional for 2-dimensional optimization in probit analysis, is easily developed for the case of 1-dimensional optimization of \$. First, for the benefit of individuals familiar with generalized linear models, the probit model can be written in the following form:

$$M^{-1}\,[\,p(\,x\,;\mu,\,\$\,)\,] = (\,x\; - \mu\,)\,\$ = x^{*}\,\mathrm{C}\,\$.$$

where, as previously, x denotes the common logarithm of dose. From this it is evident that the 1-parameter model with μ fixed can be treated as a generalized linear model with a single regressor x^{*} (=$\log_{10}$(dose)-μ), with no intercept term, and with link function M^{-1} (the probit link). As usual in probit analysis, binomial variation is assumed at a given dose, which results in a factor of $p(1\text{-}p)$ in the regression weights.

The standard approach leads to the following scheme for updating the estimate of \$:

$$[\ \$\ \text{at (I+1)th iteration}\ ] = [\ \$\ \text{at ith iteration}\ ] + \mathrm{d}_{\$}$$

where $\mathrm{d}_{\$}$ can be computed with the expression:

$$\mathrm{d}_b = \left(\frac{\P}{\P b} l(m,b)\right) \Bigg/ \left(\sum_{i=1}^{g} w_i x_i^{*2}\right)$$

with the quantities w_i, x_i, y_i defined in the following steps (recall definitions given already for d_i , r_i, and n_i):

x_i^* = $\log_{10}(d_\mathrm{i}) - \mu$ = value of “regressor” for ith treatment level, I=1,...,g;
Probit$_\mathrm{i}$ = x_i^* C \$ = predicted probit value for ith treatment level;
p_i = M(Probit$_\mathrm{i}$) = predicted response fraction at ith treatment level;
binV$_\mathrm{i}$ = p_i (1 - p_i) / n_i = binomial variance.
f_i = exp(-Probit$_\mathrm{i}^2$ / 2) / %(2B)
= weight contribution associated with probit dose-response function;
w_i = $f_i{}^2$ / binV$_\mathrm{i}$ = weight for ith treatment level;
$p_\mathrm{i}^{\mathrm{obs}}$ = r_i / n_i = observed response fraction at ith treatment level;

and

$$\frac{\P}{\P b} l(m,b) = \sum_{i=1}^{g} \; f_i \bullet x_i^* \bullet (p_i^{\mathrm{obs}} - p_i) \bullet \mathrm{binV}_i^{-1}$$

the last quantity being the partial of the 2-parameter loglikelihood with respect to \$.

\$ is not constrained to be non-negative in these computations. An argument can be made for constraining \$ to be non-negative, or greater than some small positive value such as 0.5. Adding a constraint of this sort does not appear to be technically difficult, and would probably narrow some of the confidence intervals.

Convergence criteria. All that is needed from the 1-dimensional optimization is a profile-loglikelihood value. A relative gradient criterion can be used. Convergence occurs when

$$\frac{\P}{\P b} l(m,b) \Bigg/ l(m,b) \qquad \# \; 0.00001.$$

For 2-dimensional optimization, a criterion based on relative change in parameter values is used currently.

Stabilization of parameter changes. When the starting values are too far from the optimum, the search direction indicated by the algorithm may be reasonable, while the magnitude of change in that direction may be such as to miss the optimum significantly. Improvements on the basic algorithm may involve use of the search direction, with modification of the magnitude of change in that direction, for example by use of halving or line searching (Myers, 1990, particularly Section 9.4; Seber and Wild, 1989).

For the 1-dimensional optimizations, the magnitude of parameter change ($d_{\$}$) is constrained to absolute values not exceeding 0.5. $d_{\$}$ is set to 0.5 whenever $d_{\$}$ is greater than 0.5 and $d_{\$}$ is set to -0.5 whenever $d_{\$}$ is less than -0.5. This feature eliminated some problems that occurred otherwise.

Computation of starting values for optimizations. Convergence is expected to be rapid and reliable within a sufficiently small neighborhood of the optimum. Many authors emphasize computation of starting values likely to be close to the optimum solution. In the case of probit analysis, an obvious approach for computing starting values is by a linear regression of transformed response fraction (probit transformation) against log dose. The probit transformation is not finite valued if the response fraction is 0 or 1, hence a small constant may be added or subtracted from the observed response fractions, to obtain finite probit values for use in the regression.

A starting slope value is not calculated from the data when fitting the probit function. Experience with the standard Gauss-Newton algorithm used in probit analysis has shown that numerical failures may be associated with computation of weights. Note that the weight computations involve division by the quantities $p_i(1 - p_i)$ where p_i is the predicted response fraction at the ith treatment level based on the current parameter values. Numerical failures are often related to values of one or more of the p_i that are too close to 0 or 1, so that division by zero occurs. This outcome can be prevented by setting the initial value of the slope at a small value
(a value of 0.5 is used). For a starting value for the LD50, the geometric average of test doses is used.

A starting slope value from the data is not calculated when fitting the probit function. Instead, for a starting value for the LD50, the geometric average of test doses is used. Starting values of the slope are also needed for the 1-dimensional optimizations of $\$$ (fixing μ). For most of these optimizations, an optimized value of $\$$ corresponding to a nearby value of μ is available. Otherwise, a value of 0.5 can be used.

1.5 Classification of Cases

It is proposed that whether a confidence interval can be calculated, and if so the computations to be used, will be based on the following classification (see Table A.3).

In development of this scheme, discussions with the OECD avian stat group have been very helpful, although that group has developed a somewhat different classification (report in press). For example, the avian scheme does not explicitly use the results of Silvapulle.

The conditions for cases are checked in the order that the cases are displayed in the table, so when the conditions for a given case are met, none of the higher-number cases obtain. Table A.4

indicates the computational procedures proposed for each case. Subsequent text expands upon the suggestions summarized in this table.

Table A.3. Classification of Data Cases for Purposes of Confidence Interval Calculation

Case	Description
1	("No positive dose-response association"). There is no variation in response (all animals tested in the study responded, or none responded), or the geometric mean dose is lower for animals responding than for animals not responding.
2	("Standard 2-parameter probit computations"). The Silvapulle criterion holds (i.e., one or animals responded at a dose below some other dose where one or more did not respond. The conditions defining Case 1 do not hold.
3	("No partial response fractions."). All doses tested are associated with response fractions of 0% or 100%, with the doses associated with 0% response lower than the doses associated with 100% response. One or more doses is associated with 0% response and one or more doses is associated with 100% response.
4	("One partial response fraction, first subcase"). There is a single dose associated with a partial response fraction. One or more lower test doses is associated with 0% response, and one or more higher test doses is associated with 100% response.
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).

Acknowledgement. The development of this scheme was helped by conversations with the OECD avian acute statistics group, which has developed a similar classification (report in press).

Table A.4. Classification of Data Cases for Purposes of Confidence Interval Calculation with Computational Procedures

Case	Description	Confidence interval approach	2-parameter MLE calculated	Profile likelihood procedures	
				log-likelihood supremum	**center of search region**
1	No positive dose-response association	no confidence interval computed	no	not applicable	
2	Standard 2-parameter probit computations	basic profile likelihood approach	yes	equal to loglikelihood evaluated at the MLEs	MLE for LD50
3	No partial response fractions and not Case 1.	lower bound is highest with 0% response, etc.	no	not applicable	
4	One partial response fraction, 0% response at some lower doses and 100% at some higher doses	profile loglikelihood extended by special computations	no	expression in footnote[1]	Dose associated with partial response
5	One partial response fraction, at either high test dose or low test dose	profile loglikelihood extended by special computations	no	expression in footnote[1]	Dose associated with partial response

[1] Suppose the jth dose is associated with a partial response. Then the loglikelihood supremum is

$$l_{\text{sup}} = r_j \bullet \ln\left[\frac{r_j}{n_j}\right] + (n_j - r_j) \bullet \ln\left[\frac{n_j - r_j}{n_j}\right]$$

where n_j and r_j denote the number of animals treated and the number that respond at the jth treatment level (see Section 1).

The decisions indicated in the table are as follows:

Case 1. With the stopping rules indicated for Revised UDP, this case appears to be possible only if testing is stopped at a limit dose (based on non-response for three animals tested in sequence at the dose). No methods are proposed here for cases where there is not an observable relationship between dose and response. In some cases, a binomial test may be used to establish that the LD50 is above or below the range of doses tested, but a significant binomial test requires testing of 5 or more animals and would use only the data from one test dose. Some procedures that may be applicable in this case have been developed for avian acute testing (report in press).

Case 2. Where the data allow, both probit parameters are estimated using maximum likelihood. The loglikelihood supremum is the value of the 2-parameter loglikelihood, evaluated at the MLEs.

This loglikelihood supremum is used to calculate a critical loglikelihood, which the bound values must satisfy. A search above the LD50 MLE is used to calculate an upper bound and a search below the LD50 MLE is used to calculate a lower bound.

Case 3. When there are no partial response fractions (along with other requirements of the case, as indicated in Table A.4) the profile loglikelihood approach apparently cannot be used. In this case, it seems that any dose within the interval bounded by the highest dose with no response, and the lowest dose with 100% response, would be equally valid as an estimate of the LD50. Simulations suggest that these two doses will perform acceptably when used as confidence bounds. Graphs of the profile loglikelihood indicate discontinuities at those doses, so that the profile loglikelihood approach cannot be implemented in a straightforward manner.

Case 4. When there is a single partial response (along with other requirements for the case, as indicated), the profile loglikelihood can be applied using special computations as described in Section 1. It is proposed that, when searching numerically for a bound, the dose with partial response can be used to define the search interval.

Case 5. This is an infrequent case which occurs mainly if the LD50 is close to a bound.

1.6 References

Barndorff-Nielsen, O.E. 1991. Likelihood theory. Chapter 10 in D.V. Hinkley, N. Reid, and E.J. Snell (eds) *Statistical Theory and Modelling*. Chapman and Hall.

Bickel, E.J. and K. A. Doksum. 2001. *Mathematical Statistics: basic ideas and selected topics*. Volume 1. (2nd ed.) Prentice Hall.

Finney, D.J. 1971. *Probit Analysis*. 3rd ed. Cambridge U. Press.

McCullagh, P., and J.A. Nelder. 1989. *Generalized linear models*. (2nd ed.) Chapman & Hall/CRC.

Myers, R.H. 1990. *Classical and modern regression with applications*. (2nd ed.) Duxbury Press.

Seber, G.A.F., and Wild, C.J. 1989. *Nonlinear regression*. John Wiley and Sons.

Silvapulle, M.J. 1981. On the existence of a maximum likelihood estimators for the binomial response model. J. Royal Statist. Soc. Series B 43(3):310-313.

Appendix B

Tables of Simulation Results

Table B.1. Percentages of Cases 1-5 among Simulated Studies

Scenario#	LD50	slope	initial test dose	% Case 1 + Case 5	% Case 2	% Case 3	% Case 4
(I) Scenarios with initial test dose 175 units							
1	1.5	8.33	175	21.3	0.0	78.7	0.0
2		4	175	53.1	3.0	42.6	1.3
3		2	175	41.6	31.0	18.5	8.9
4		0.8	175	19.7	61.6	6.0	12.7
5		0.5	175	11.5	67.7	5.7	15.1
6	2.5	8.33	175	0.0	0.0	99.2	0.8
7		4	175	10.8	6.3	64.7	18.2
8		2	175	13.9	38.5	21.9	25.8
9		0.8	175	10.6	66.3	6.7	16.5
10		0.5	175	9.3	70.4	5.1	15.2
11	20	8.33	175	0.0	0.0	35.3	64.7
12		4	175	0.0	9.3	24.9	65.8
13		2	175	0.0	40.6	14.8	44.6
14		0.8	175	0.2	61.9	7.6	30.3
15		0.5	175	1.5	61.7	7.2	29.7
16	50	8.33	175	0.0	0.0	29.8	70.2
17		4	175	0.0	7.3	24.0	68.7
18		2	175	0.0	37.2	12.4	50.4
19		0.8	175	0.0	54.9	8.4	36.7
20		0.5	175	0.3	57.5	7.0	35.2
21	150	8.33	175	0.0	0.0	36.7	63.3
22		4	175	0.0	4.1	26.6	69.3
23		2	175	0.0	26.1	15.8	58.1
24		0.8	175	0.0	48.5	9.8	41.7
25		0.5	175	0.6	56.6	8.0	34.9
26	600	8.33	175	0.0	0.0	30.3	69.7
27		4	175	0.0	6.7	22.9	70.4
28		2	175	0.0	32.6	12.7	54.7
29		0.8	175	0.6	54.3	9.0	36.1
30		0.5	175	1.9	58.9	8.6	30.6
31	1500	8.33	175	0.0	0.0	39.9	60.1
32		4	175	0.2	9.3	24.8	65.8
33		2	175	1.2	43.4	13.5	41.9
34		0.8	175	4.4	59.8	6.5	29.3
35		0.5	175	6.0	62.1	5.8	26.0
36	3000	8.33	175	9.5	1.1	82.4	7.0

Scenario#	LD50	slope	initial test dose	% Case 1 + Case 5	% Case 2	% Case 3	% Case 4
37		4	175	21.0	25.4	30.0	23.5
38		2	175	14.7	52.4	11.3	21.6
39		0.8	175	11.2	62.9	6.4	19.5
40		0.5	175	11.2	60.1	5.2	23.5
41	3500	8.33	175	27.4	1.0	70.5	1.1
42		4	175	36.1	24.9	28.0	11.1
43		2	175	22.4	50.9	9.5	17.2
44		0.8	175	12.1	62.6	6.4	18.9
45		0.5	175	11.0	60.0	6.8	22.3
(ii) Scenarios with initial test dose varied							
46	1.5	8.33	1.1	0.0	2.1	66.0	31.9
47			1.5	1.2	9.2	22.0	67.6
48			1.9	8.0	9.3	43.1	39.6
49		4	1.5	4.2	27.6	16.1	52.1
50			2.4	18.6	27.8	23.9	29.7
51		2	1.5	9.1	40.6	12.3	38.0
52			4	30.9	39.0	14.1	16.0
53		0.8	1.5	15.5	52.5	6.2	25.9
54			16.9	19.5	58.7	6.5	15.2
55		0.5	1.5	19.3	50.0	6.7	24.0
56			72.3	8.2	67.4	5.8	18.6
57	2.5	8.33	1.8	0.0	0.1	67.6	32.3
58			2.5	0.0	0.0	26.1	73.9
59			3.1	0.0	0.0	50.1	49.9
60		4	1.2	0.0	10.1	33.4	56.5
61			2.5	0.7	8.2	22.6	68.4
62			4.1	6.5	8.9	43.3	41.3
63		2	2.5	3.1	38.4	14.1	44.5
64			6.6	1.5	40.0	13.6	44.9
65		0.8	2.5	11.8	53.3	7.0	28.0
66			28.2	6.8	60.4	7.5	25.3
67		0.5	2.5	14.0	54.3	6.4	25.4
68			120.5	7.1	67.3	5.9	19.6
69	20	8.33	14	0.0	0.2	74.1	25.7
70			20	0.0	0.0	25.7	74.3
71			25.2	0.0	0.0	50.0	50.0
72		4	9.6	0.0	9.5	34.0	56.5
73			20	0.0	5.0	21.7	73.3
74			32.5	0.0	10.8	34.3	54.9

Scenario#	LD50	slope	initial test dose	% Case 1 + Case 5	% Case 2	% Case 3	% Case 4
75		2	4.6	0.0	41.1	14.6	44.3
76			20	0.0	28.1	16.6	55.3
77			52.7	0.0	32.3	14.3	53.4
78		0.8	20	0.1	51.2	8.9	39.8
79			225.4	0.1	62.4	7.8	29.7
80		0.5	20	0.9	59.5	8.2	31.4
81			964.4	1.5	71.6	6.3	20.6
82	50	8.33	35.1	0.0	0.0	73.9	26.1
83			50	0.0	0.0	22.6	77.4
84			63.1	0.0	0.0	50.8	49.2
85		4	23.9	0.0	9.2	36.1	54.7
86			50	0.0	3.3	20.9	75.8
87			81.2	0.0	8.8	34.8	56.4
88		2	11.4	0.0	35.6	15.0	49.4
89			50	0.0	27.4	14.2	58.4
90			131.8	0.0	32.1	13.5	54.4
91		0.8	1.3	0.0	68.1	7.6	24.3
92			50	0.0	51.8	8.0	40.2
93			563.6	0.0	58.7	8.6	32.7
94		0.5	50	0.8	57.4	8.2	33.5
95			2411.1	1.5	69.6	7.2	21.8
96	150	8.33	105.3	0.0	0.0	71.8	28.2
97			150	0.0	0.0	24.6	75.4
98			189.3	0.0	0.0	50.0	50.0
99		4	71.7	0.0	9.5	33.0	57.5
100			150	0.0	4.9	21.4	73.7
101			243.5	0.0	9.3	34.7	56.0
102		2	34.3	0.0	36.0	14.5	49.5
103			150	0.0	26.7	16.8	56.5
104			395.3	0.0	32.0	13.6	54.4
105		0.8	3.8	0.2	70.0	5.4	24.5
106			150	0.0	51.5	9.1	39.4
107			1690.9	0.3	62.0	8.0	29.7
108		0.5	150	0.7	55.7	8.2	35.4
109	600	8.33	421	0.0	0.1	72.7	27.2
110			600	0.0	0.0	26.9	73.1
111			757.2	0.0	0.1	51.4	48.5
112		4	286.9	0.0	11.4	33.7	54.9
113			600	0.0	4.3	25.3	70.4

Scenario#	LD50	slope	initial test dose	% Case 1 + Case 5	% Case 2	% Case 3	% Case 4
114			974	0.0	13.2	35.8	51.0
115		2	137.2	0.0	36.8	14.8	48.4
116			600	0.1	26.7	16.3	56.9
117			1581.1	0.4	31.8	14.1	53.7
118		0.8	15	0.1	69.6	7.1	23.2
119			600	1.2	52.9	8.4	37.5
120		0.5	1.6	1.5	75.5	5.1	17.9
121			600	3.0	59.6	6.4	31.0
122	1500	8.33	1052.5	0.0	0.0	72.9	27.1
123			1500	0.0	0.0	23.4	76.6
124			1892.9	0.0	0.0	52.4	47.6
125		4	717.3	0.0	7.5	34.2	58.3
126			1500	0.2	5.1	23.6	71.2
127			2435	0.0	9.4	34.6	56.0
128		2	343	3.8	39.8	17.2	39.2
129			1500	3.0	30.3	14.2	52.6
130			3952.8	0.5	37.2	13.9	48.4
131		0.8	37.5	4.7	69.0	6.2	20.1
132			1500	11.0	51.8	7.4	29.8
133		0.5	4.1	5.2	74.6	5.2	15.0
134			1500	15.1	52.4	6.6	25.9
135	3000	8.33	2105.1	5.4	2.9	66.4	25.3
136			3000	0.8	4.9	23.2	71.1
137			3785.8	0.2	1.5	52.7	45.6
138		4	1434.6	27.3	14.9	39.7	18.1
139			3000	3.0	24.8	20.1	52.1
140			4750	0.7	17.7	35.1	46.6
141		2	686	11.5	46.3	11.9	30.3
142			3000	8.5	40.5	11.8	39.2
143		0.8	75	7.9	67.5	5.4	19.2
144			3000	14.0	52.7	4.8	28.5
145		0.5	8.2	5.5	76.4	4.5	13.6
146			3000	18.2	52.9	6.9	22.0
147	3500	8.33	2455.9	17.8	6.6	53.3	22.3
148			3500	1.9	13.4	19.8	64.9
149			4416.8	0.0	4.8	50.5	44.7
150		4	1673.7	37.8	18.0	28.3	15.9
151			3500	4.7	30.4	16.0	48.9
152		2	800.4	13.9	50.0	9.0	27.1

Scenario#	LD50	slope	initial test dose	% Case 1 + Case 5	% Case 2	% Case 3	% Case 4
153			3500	8.1	43.4	11.2	37.3
154		0.8	87.5	9.2	66.8	6.0	18.0
155			3500	15.5	52.4	5.8	26.2
156		0.5	9.6	13.3	69.3	5.1	12.3
157			3500	18.1	54.2	5.2	22.4

Table B.2. Performance Measures PM1-PM4 (defined in Section 2.5).

Scenario	LD50	slope	initial	PM1(%)	PM2(%)	PM3	PM4(%)
#			test				
			dose				
(I) Scenarios with initial test dose of 175 units							
1	1.5	8.33	175	100.0	100.0	5.5	302
2		4	175	98.4	99.6	5.5	302
3		2	175	76.4	93.8	10.5	449
4		0.8	175	53.8	87.2	>1000	3033
5		0.5	175	45.2	79.6	>1000	>100000
6	2.5	8.33	175	100.0	100.0	5.5	181
7		4	175	99.6	99.9	5.5	181
8		2	175	89.7	96.5	7.0	275
9		0.8	175	58.2	88.1	>1000	2167
10		0.5	175	46.6	80.9	>1000	>100000
11	20	8.33	175	100.0	96.0	4.2	178
12		4	175	99.1	92.7	4.2	178
13		2	175	88.2	89.0	8.8	213
14		0.8	175	58.0	77.4	260.5	3425
15		0.5	175	52.5	73.2	>1000	5029
16	50	8.33	175	100.0	95.3	4.0	118
17		4	175	97.0	90.7	4.4	185
18		2	175	75.2	88.8	11.1	269
19		0.8	175	56.8	85.6	89.4	2012
20		0.5	175	52.2	81.8	>1000	4332
21	150	8.33	175	100.0	97.8	24.5	457
22		4	175	95.9	93.9	24.5	457
23		2	175	74.1	88.7	24.5	457
24		0.8	175	56.3	80.6	24.5	1250
25		0.5	175	50.0	79.1	>1000	>100000
26	600	8.33	175	100.0	93.8	4.0	191
27		4	175	96.9	89.2	4.2	191
28		2	175	77.8	89.0	10.5	224
29		0.8	175	55.3	84.0	63.3	4092
30		0.5	175	48.2	81.0	>1000	>100000
31	1500	8.33	175	100.0	97.1	4.1	135
32		4	175	98.8	93.0	4.1	214
33		2	175	82.6	89.0	10.3	247
34		0.8	175	51.7	79.8	>1000	>100000
35		0.5	175	44.7	76.9	>1000	>100000
36	3000	8.33	175	99.8	100.0	2.9	108

Scenario #	LD50	slope	initial test dose	PM1(%)	PM2(%)	PM3	PM4(%)
37		4	175	93.4	98.4	3.6	108
38		2	175	73.0	93.9	14.2	574
39		0.8	175	46.6	81.2	>1000	>100000
40		0.5	175	43.9	75.3	>1000	>100000
41	3500	8.33	175	99.7	100.0	2.9	93
42		4	175	90.2	99.5	3.6	93
43		2	175	64.0	94.6	108.9	1296
44		0.8	175	48.0	83.0	>1000	>100000
45		0.5	175	45.3	75.5	>1000	>100000

(ii) Scenarios with initial test dose varied

46	1.5	8.33	1.1	97.9	99.9	3.2	159
47			1.5	91.9	100.0	5.7	216
48			1.9	93.3	98.8	9.2	332
49		4	1.5	73.0	99.1	12.4	441
50			2.4	74.5	98.7	14.4	510
51		2	1.5	57.2	94.1	12.4	441
52			4	59.6	97.1	16.5	702
53		0.8	1.5	42.8	90.8	>1000	>100000
54			16.9	40.5	81.6	>1000	>100000
55		0.5	1.5	43.7	86.2	>1000	>100000
56			72.3	46.9	74.7	>1000	>100000
57	2.5	8.33	1.8	99.9	99.9	3.2	156
58			2.5	100.0	100.0	15.6	329
59			3.1	100.0	99.2	3.2	268
60		4	1.2	90.0	97.3	4.4	192
61			2.5	92.2	99.4	15.6	329
62			4.1	94.0	99.6	5.8	224
63		2	2.5	63.4	96.3	19.9	532
64			6.6	63.6	94.4	24.5	1033
65		0.8	2.5	44.4	88.3	>1000	>100000
66			28.2	52.6	79.5	>1000	4415
67		0.5	2.5	42.5	87.0	>1000	>100000
68			120.5	46.8	77.4	>1000	>100000
69	20	8.33	14	99.8	100.0	3.2	151
70			20	100.0	100.0	24.5	391
71			25.2	100.0	99.3	3.2	272
72		4	9.6	90.9	97.7	4.4	213
73			20	95.0	98.9	24.5	391

Scenario #	LD50	slope	initial test dose	PM1(%)	PM2(%)	PM3	PM4(%)
74			32.5	89.2	99.0	24.5	295
75		2	4.6	80.7	87.7	10.8	317
76			20	72.6	93.7	24.5	575
77			52.7	74.4	90.6	24.5	479
78		0.8	20	52.7	85.2	63.2	2288
79			225.4	63.6	76.6	70.2	1125
80		0.5	20	47.0	80.8	>1000	>100000
81			964.4	56.7	77.3	>1000	4874
82	50	8.33	35.1	100.0	99.9	3.2	152
83			50	100.0	100.0	24.5	391
84			63.1	100.0	99.1	3.2	86
85		4	23.9	91.5	96.8	4.4	183
86			50	96.7	99.1	24.5	391
87			81.2	91.2	98.7	24.5	295
88		2	11.4	81.5	89.1	10.9	282
89			50	72.7	91.0	24.5	575
90			131.8	74.0	90.7	24.5	479
91		0.8	1.3	72.8	77.0	81.4	2301
92			50	54.4	84.0	63.3	1238
93			563.6	66.9	73.9	33.2	973
94		0.5	50	48.7	79.3	>1000	>100000
95			2411.1	58.9	75.8	>1000	4121
96	150	8.33	105.3	100.0	99.9	3.2	152
97			150	100.0	100.0	24.5	391
98			189.3	100.0	99.7	3.2	273
99		4	71.7	90.7	97.1	4.4	251
100			150	95.1	98.9	24.5	391
101			243.5	90.8	98.9	24.5	295
102		2	34.3	83.0	91.4	8.8	272
103			150	73.5	92.0	24.5	575
104			395.3	72.2	90.7	24.5	479
105		0.8	3.8	70.0	78.6	120.2	3826
106			150	53.3	84.0	64.5	1238
107			1690.9	62.7	76.3	75.8	1139
108		0.5	150	50.5	80.0	>1000	25569
109	600	8.33	421	100.0	99.9	3.2	152
110			600	100.0	100.0	24.5	391
111			757.2	99.9	99.3	3.2	86
112		4	286.9	92.5	97.0	4.9	207

Scenario #	LD50	slope	initial test dose	PM1(%)	PM2(%)	PM3	PM4(%)
113			600	95.7	98.7	24.5	391
114			974	90.7	99.2	4.8	219
115		2	137.2	80.8	91.1	9.1	281
116			600	73.3	92.6	24.5	575
117			1581.1	74.9	91.7	24.5	266
118		0.8	15	65.1	81.2	183.2	3419
119			600	50.8	84.4	>1000	1509
120		0.5	1.6	62.2	78.7	>1000	>100000
121			600	45.5	82.5	>1000	>100000
122	1500	8.33	1052.5	100.0	100.0	4.8	263
123			1500	100.0	100.0	25.6	405
124			1892.9	100.0	99.3	3.2	86
125		4	717.3	92.6	98.1	4.0	166
126			1500	95.0	99.6	25.6	405
127			2435	90.6	98.9	10.7	295
128		2	343	74.6	94.0	6.7	261
129			1500	72.7	96.1	27.0	617
130			3952.8	69.2	87.7	10.5	358
131		0.8	37.5	59.3	84.0	>1000	63283
132			1500	46.0	90.8	>1000	>100000
133		0.5	4.1	51.7	78.5	>1000	>100000
134			1500	45.4	87.3	>1000	>100000
135	3000	8.33	2105.1	98.4	99.9	2.4	96
136			3000	95.6	100.0	7.1	225
137			3785.8	98.7	99.1	3.2	86
138		4	1434.6	89.3	99.9	3.5	119
139			3000	77.0	99.3	13.7	256
140			4750	82.8	99.1	4.4	137
141		2	686	58.6	93.2	24.5	231
142			3000	57.2	95.0	13.7	256
143		0.8	75	51.8	82.9	>1000	>100000
144			3000	42.2	90.1	>1000	>100000
145		0.5	8.2	52.3	79.9	>1000	>100000
146			3000	42.1	85.7	>1000	>100000
147	3500	8.33	2455.9	94.7	99.3	2.0	73
148			3500	87.4	100.0	5.2	172
149			4416.8	95.3	99.4	3.2	86
150		4	1673.7	83.4	99.9	9.8	254
151			3500	69.7	99.4	11.8	229

Scenario #	LD50	slope	initial test dose	PM1(%)	PM2(%)	PM3	PM4(%)
152		2	800.4	53.7	94.6	864.9	23232
153			3500	54.5	96.1	24.5	229
154		0.8	87.5	56.3	80.4	>1000	>100000
155			3500	40.8	89.8	>1000	>100000
156		0.5	9.6	46.6	81.1	>1000	>100000
157			3500	39.7	86.6	>1000	>100000