FUNDAMENTAL AND APPLIED TOXICOLOGY 5, 151–157 (1985)

An Up-and-Down Procedure for Acute Toxicity Testing

ROBERT D. BRUCE

The Procter & Gamble Company, P.O. Box 39175, Cincinnati, Ohio 45247

An Up-and-Down Procedure for Acute Toxicity Testing. BRUCE, R. D. (1985). *Fundam. Appl. Toxicol.* **5**, 151–157. An up-and down method for acute toxicity (LD50) testing has been developed and statistically evaluated. Compared with the "classical" procedure, this method permits a major reduction in the number of animals used. In the up-and-down procedure, animals are dosed one at a time. If an animal survives, the dose for the next animal is increased; if it dies, the dose is decreased. A survey of 48 acute toxicity tests in rats showed that the great majority of the animals that ultimately died did so within 1 or 2 days. Because of this, it suffices to observe each animal for 1 or 2 days before dosing the next animal. It is recommended, however, that surviving animals be monitored for delayed death for a total of 7 days. The procedure for estimating the LD50 takes into account all deaths, and may be performed using widely available computer program packages. Testing in females alone is recommended, based on the observation that they were generally more sensitive in the survey of 48 studies; selective follow-up in males may sometimes be indicated. The procedure has been tested, by simulation, on 10 of the survey studies. It produced excellent agreement with the original studies. The 95% confidence interval for the LD50 averaged ±32% by the up-and-down method, compared with ±15% for conventional studies using 40 to 50 animals. The up-and-down procedure will require only 6 to 10 animals, provided that the initial estimate of the LD50 is within a factor of 2 of the true LD50. The method cannot be recommended for testing materials where deaths beyond 2 days postdosing are the rule.

The up-and-down method of experimentation is an adaptive procedure for conducting dose–response experiments having a yes–no endpoint. Using this strategy for acute toxicity testing, animals are dosed one at a time, starting the first animal at the toxicologist's best estimate of the LD50. If this animal survives, then the next animal receives a higher dose, while if the first animal dies, the next animal receives a lower dose. Doses are usually adjusted by a constant multiplicative factor, for example 1.3. The dose for each successive animal is adjusted up or down depending upon the outcome for the previous animal. It can be seen that this method of experimentation causes the doses to be rapidly adjusted toward the LD50 and then to be maintained in the region of the LD50. Thus this procedure concentrates experimental effort in the most relevant region and, as a result, uses animals in a very efficient manner. For this reason, the method is a logical choice if one wishes to minimize the number of animals required to estimate the LD50.

The development of this method was based upon two investigations. The first of these, a historical review of a large number of conventional acute toxicity studies, was used as a basis for making recommendations about the length of time between successive animals, the sex of the animals to be tested, and the dose multiplier. The second investigation, a set of simulation experiments, provided a test of the procedure using data drawn from historical studies. Both the accuracy and the precision of the up-and-down method were evaluated in this latter study.

The up-and-down method was originally developed during World War II for the purpose of testing the sensitivity of explosives and was published by Dixon and Mood (1948). In this article, the authors foresaw

0272-0590/85 $3.00

the value of their method in testing the lethality of insecticides and in the general areas of biological and medical research. Later publications by Brownlee *et al.* (1953) and Dixon (1965) discussed the up-and-down method with small numbers of subjects. The textbook of Dixon and Massey (1969) includes a fairly comprehensive description of the method.

EXPERIMENTAL

Historical survey. This part of the investigation was based upon a review of 48 conventional acute toxicity tests covering a variety of chemicals and finished product formulations having moderate oral toxicity. The materials tested included laundry detergents, hard surface cleaning products, dishwashing detergents, antiperspirants, shampoos, analgesics, and ingredients used in the manufacture of these products. These tests typically included 40 to 50 animals (Sprague–Dawley rats) in groups of 10 animals, 5 of each sex. Animals were observed at least daily for 14 days after dosing. The data concerning time until death were tabulated with the SAS system (SAS Institute, Inc., 1982a). For each test, the dose–response patterns for males and females were compared by the Mantel–Haenszel procedure (Mantel, 1963). A probit analysis (Finney, 1971) was performed on each study, using the pooled data for males and females, by the PROBIT procedure of the SAS system. This analysis gave an LD50 value and an estimate of σ (the reciprocal of the slope of the probit–log dose plot) for each of 42 studies; for 6 of the studies, the data did not permit probit analysis. Probit analyses were also performed separately by sex where the data were suitable.

Simulation experiments. For 10 of the historical studies, simulations of the up-and-down method were conducted. For each study, individual animal records were available and included time of death. The simulation was conducted by selecting, at random, an animal from the lowest dose group of the study. Successive animals were randomly selected from higher or lower dose groups depending upon whether the previous animal lived or died within a specified time period (1 or 2 days). In a few instances an animal selected from the highest dose group survived for the specified time period and hence selection of the next animal from a still higher dose group was indicated. In these cases, it was assumed that 100% mortality would have occurred at the higher dose. The process was continued until five animals had been tested after reversal of the initial outcome; for example if the experiment began with an animal that survived, five animals were tested counting from the first death.

The data from the simulated up-and-down studies were analyzed by the method of maximum likelihood (Dixon, 1965; Finney, 1971) using the SAS procedure NLIN (SAS Institute, Inc., 1982b). The model used had parameters μ and σ, where μ is the log of the LD50 and σ is the standard deviation of the tolerance distribution. Based upon the results of the historical survey, σ was fixed at values of 0.06, 0.12, or 0.24. All deaths, whether immediate or delayed, were considered as deaths for the purpose of the maximum likelihood analysis. Dixon (1965) shows how the likelihood function can be written.

RESULTS

Among the 48 historical studies, a total of 2070 animals were tested and, of these, 990 died during the 14-day observation period. Figure 1 shows the distribution of days until death. From this figure we can see that the overwhelming majority of deaths occurred in the first few days and that only four deaths (0.4%) occurred more than 7 days after dosing. Delayed deaths are defined as those deaths occurring more than 1 or 2 days after dosing. Figure 2 shows that relatively few studies had a high percentage of delayed deaths.

The Mantel–Haenszel procedure revealed that males had higher LD50 values than females in a majority of cases. Figure 3 shows the results of this comparison expressed in terms of the direction and significance of the sex difference. For only 3 of the 48 studies was there evidence of lower LD50 values

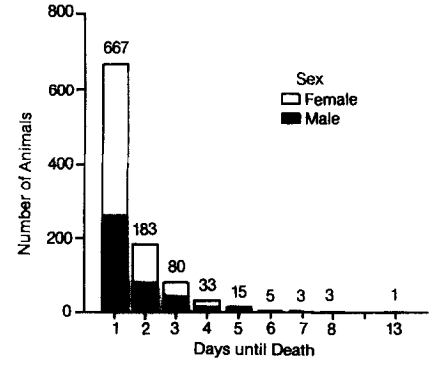

FIG. 1. Distribution of days until death for rats in 48 historical studies.

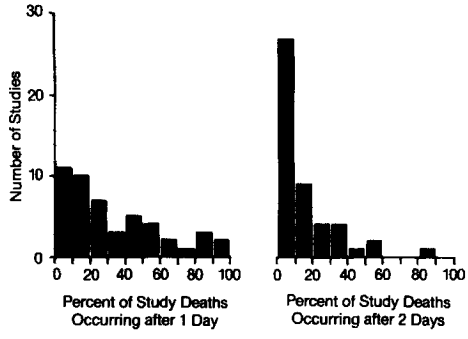

FIG. 2. Number of studies with delayed deaths among 48 historical studies.

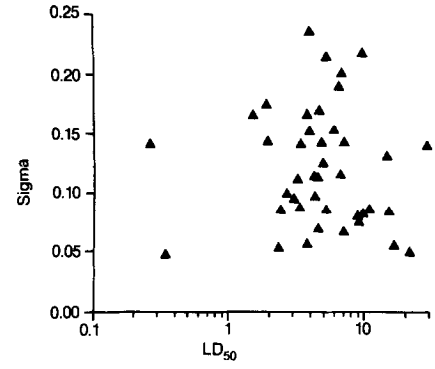

FIG. 4. The relation between σ and the LD50 in 42 historical studies.

among males, and none of these tests approached statistical significance. By contrast, in 13 of the studies, the males had significantly higher LD50 values than the females (risk <5%). While the Mantel–Haenszel procedure could be conducted for all 48 studies, only 16 of the studies would permit probit analysis on both males and females. For these studies, the LD50 value for males averaged 29% higher than that for females.

For the 42 studies where data (sexes combined) permitted probit analysis, the absence of any relation between σ (the reciprocal of the slope of the probit vs log dose plot), and the LD50 value is shown in Fig. 4. The average value of σ was 0.121 with a range from 0.048 to 0.237. Figure 5 illustrates the data generated when the up-and-down procedure was simulated by random sampling from Historical Study No. 19. In this simulation, the survival/mortality seen within the first day after dosing each animal was used as the basis for deciding whether the next animal should be selected from a higher or lower dose group. The results of performing similar simulations on 10 historical studies are shown in Table 1. The number of animals ranged from six to nine. As was noted earlier, the method of estimating the LD50 requires that a value of sigma (σ) be supplied. The effect of varying the value of σ upon the resulting logarithm of the LD50 is illustrated in Table 1 for values of σ = 0.06, 0.12, and 0.24. The changes in log LD50 values are seen to be minor. In the worst case, Study 16, the LD50 using σ = 0.24 is 9% below the LD50 obtained for σ = 0.12. Similarly, the final three columns of Table 1 show the

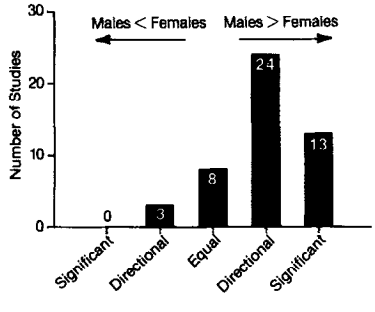

FIG. 3. The effect of sex of rats upon the LD50 in 48 historical studies.

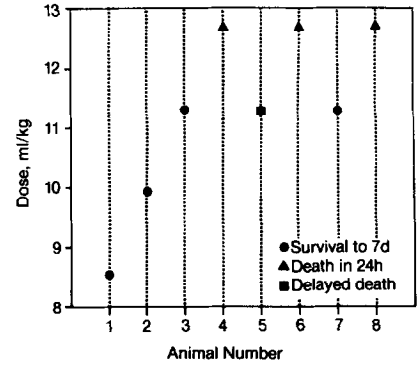

FIG. 5. An example of a simulated up-and-down study; data drawn from Study No. 19 using a 1-day delay between successive animals.

TABLE 1

SIMULATION OF UP-AND-DOWN PROCEDURE BASED ON 10 HISTORICAL STUDIES[a]

Study	Full study: Log LD_{50}	Full study: SE	n	Simulated up-and-down study: Log LD_{50} $\sigma = 0.06$	Log LD_{50} $\sigma = 0.12$	Log LD_{50} $\sigma = 0.24$	SE $\sigma = 0.06$	SE $\sigma = 0.12$	SE $\sigma = 0.24$
6	−0.462	0.016	7	−0.505	−0.518	−0.563	0.037	0.079	0.134
3	0.182	0.039	8	0.316	0.319	0.344	0.041	0.064	0.115
14	0.382	0.013	6	0.337	0.337	0.337	0.032	0.062	0.123
15	0.854	0.034	7	0.934	0.936	0.950	0.038	0.065	0.120
16	0.832	0.027	8	0.863	0.835	0.793	0.041	0.064	0.114
17	0.639	0.031	7	0.612	0.605	0.609	0.043	0.069	0.124
18	0.586	0.014	7	0.595	0.579	0.556	0.036	0.061	0.117
19	1.050	0.020	8	1.055	1.052	1.051	0.030	0.055	0.107
26	−0.362	0.024	6	−0.421	−0.429	−0.489	0.071	0.107	0.178
48	0.539	0.027	9	0.609	0.597	0.591	0.045	0.068	0.114

[a] 1 day delay between animals.

effect that varying σ has upon the standard error of the logarithm of the LD50. Here, the standard errors are seen to be almost directly proportional to the assumed value of σ. For the central value of $\sigma = 0.12$, the average standard error is 0.069, in logarithmic units. The LD50s obtained by simulation of the up-and-down procedure are compared with the LD50s using all of the study data in Fig. 6. This figure shows results for both a 1-day delay between animals and for a 2-day delay; the latter results are shown only where they differed from those obtained using a 1-day delay. It can be seen that, for either delay period, the up-and-down results are in excellent agreement with the full study.

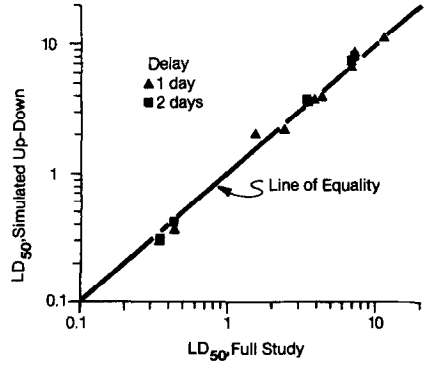

FIG. 6. Comparison of the LD50 for the simulated up-and-down procedure with the LD50 obtained from the full study; based on 10 studies, $\sigma = 0.12$.

DISCUSSION

From the distribution of days until death it can be concluded that, at least for the types of materials tested in these studies, death either occurs within the first few days after dosing or the animal recovers. Deaths after 7 days were so unusual as to lead to a recommended total observation period of 7 days. And the fact that most animals dying did so in the first day indicates that it should usually suffice to observe an animal for a single day before deciding upon the dose level for the next animal. Using the actual outcome after 1 day (dead or alive) augmented by the observer's judgment (animal appears moribund after 1 day) should lead to successful choice of the next dose in most cases. The estimation procedure used will allow for "errors" in choice of the next dose, since it uses the actual pattern of deaths at the end of the full observation period. However, the method does require that the final results show a logical progression from survival at lower doses to death at higher doses.

In the case of materials which are quite slow-acting, it is possible that the dose would be continuously increased. Such a material could be tested successfully by the up-and-down method only if the time between successive animals were increased. Figure 2 shows that, for this sample of historical studies, very few materials produced a preponderance of delayed deaths. Nevertheless, it is recommended that any protocol implementing this method of testing should include a provision for termination in the case where most or all animals die late in the observation period. If the rule adopted is to lower the dose if the earlier animal dies or is moribund after the initial observation period, then another possible difficulty could be encountered. An animal appearing moribund could survive leading to a lower dose for the next animal when, ideally, the dose should have been increased. If this happened every time the lethal dose was approached, the result could be an experiment in which no animals died and it would not be possible to estimate an LD50. While this latter result seems unlikely, provision could be made in the protocol to terminate the study in this case.

Only 3 of the 48 historical studies showed male animals to have lower LD50 values than females. This suggests that, for materials like those studied here, it generally would be conservative to use female animals alone. Should there be a desire to check the possibility of a lower LD50 among males, this could be done in a followup study in which a single group of males is dosed at the LD50 established for females. If more than half of such a group died, it would suggest a lower LD50 value for the males and the need for further experimentation. Another possibility would be to run the up-and-down procedure concurrently among males. It is not recommended that the up-and-down procedure be applied to a group of mixed sex unless it has already been established that, for the type of material being tested, sex differences are unlikely. Where there are sex differences similar to those reported above, an up-and-down experiment on a group of mixed sex would be likely to produce erratic results. Up-and-down strategies for dosing animals two (or more) at a time have been developed (Wetherill, 1963; Tsutakawa, 1967; Hsi, 1969), but application of such a method to a group of mixed sex (e.g., one male and one female in each experimental unit) would seem to offer no advantage over simply running a separate up-and-down experiment for each sex.

The work of Dixon (1965) recognized that an experiment based on a small number of animals would not permit internal estimation of the steepness of the dose–response curve or its related parameter σ, and he implied that σ might need to be based upon historical data. Dixon also recommended that σ be chosen as the spacing between doses, on a logarithmic scale. He also showed that the computed LD50 was relatively insensitive to changes in the value of σ, while the standard error of the logarithm of the LD50 was shown to be directly proportional to σ. Both of these results are confirmed by the present work (see Table 1). The value of σ recommended for use is 0.12, about the average of the historical data reported here. Since the standard error of the logarithm of the LD50 value depends upon this assumed value of σ, then so also will estimates of the confidence interval for the LD50. An approximate 95% confidence interval for the logarithm of the LD50 can be obtained from

$$\log \text{LD50} \pm 1.96 \text{ (SE of log LD50)},$$

and a corresponding 95% confidence interval for the LD50 may be obtained by taking antilogarithms of the values given by this expression. Applying this expression to the average standard error for the 10 simulated up-and-down study results (0.069) gives an estimated 95% confidence interval of

$$\log \text{LD50} \pm 0.135,$$

which, upon taking antilogarithms, gives an interval from 0.73 LD50 to 1.37 LD50 (−27% to +37% of the LD50). The historical value of σ, combined with Dixon's recommenda-

TABLE 2

DATA AND ANALYSES FOR HISTORICAL STUDY 17

(a) Data

Dose (ml/kg)	No. of deaths/animals tested			Day of death	
	Male	Female	Total	Male	Female
1.54	0/5	0/5	0/10	—	—
2.94	0/5	1/5	1/10	—	2
4.34	0/5	4/5	4/10	—	1, 1, 1, 2
5.74	4/5	5/5	9/10	1, 1, 1, 1	1, 1, 1, 1, 1

(b) Probit analysis (based on Total column above)

Log (LD50) = 0.639

LD50 = 4.35

σ = 0.115

(c) Mantel–Haenszel test (males vs females)

χ^2 = 5.357

p value = 0.021

tion, suggests that the spacing between successive doses in the up-and-down procedure should be around 0.12 logarithmic units which corresponds to a dose multiplier of about 1.3.

Those who wish to adopt this procedure may wish to consider conducting their own historical survey to confirm, in particular, the extent of delayed deaths, the effect of sex and the typical value of σ. Such a historical survey would follow the steps outlined previously. Table 2 shows the data and results of these analyses for Historical Study No. 17. The prospective user would repeat these steps for a number of representative studies from historical records and then summarize the results as illustrated earlier. The stopping rule (five animals after reversal of initial outcome) may be altered to increase or decrease the number of animals used and the relative precision of the LD50 value. A strategy has also been published (Cochran and Davis, 1965) in which the spacing between doses is altered as the experiment proceeds. For the present purpose, this method was felt to offer relatively little advantage to justify the added complexity.

This method offers the potential of substantial savings in numbers of animals although the estimated LD50 values will be less precise than those obtained from larger experiments. When other sources of error, such as laboratory to laboratory variation, differences in strain, and interspecies effects are considered, however, the precision obtained with the up-and-down method may be sufficient for most practical purposes. At the same time, it must be recognized that this method may be inappropriate for materials typically producing death 2 or more days after administration, in cases where government regulations dictate the test method to be used, and where high levels of precision are required.

REFERENCES

BROWNLEE, K. A., HODGES, J. L., JR., AND ROSENBLATT, M. (1953). The up-and-down method with small samples. *J. Amer. Stat. Assoc.* **48**, 262–277.

COCHRAN, W. G., AND DAVIS, M. (1965). The Robbins-Munro method for estimating the median lethal dose. *J. R. Stat. Soc. Ser. B* **27**, 28–44.

DIXON, W. J. (1965). The up-and-down method for small samples. *J. Amer. Stat. Assoc.* **60**, 967–978.

DIXON, W. J., AND MASSEY, F. J., JR. (1969). *Introduction to Statistical Analysis*, 3rd ed. McGraw-Hill, New York.

DIXON, W. J., AND MOOD, A. M. (1948). A method for obtaining and analyzing sensitivity data. *J. Amer. Stat. Assoc.* **43**, 109–126.

FINNEY, D. J. (1971). *Probit Analysis*, 3rd ed. Cambridge Univ. Press, Cambridge, England.

HSI, B. P. (1969). The multiple sample up-and-down method in bioassay. *J. Amer. Stat. Assoc.* **64**, 147–162.

MANTEL, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *J. Amer. Stat. Assoc.* **58**, 690–700.

SAS Institute Inc. (1982a). *SAS User's Guide: Basics.* Cary, N.C.

SAS Institute Inc. (1982b). *SAS User's Guide: Statistics.* Cary, N.C.

TSUTAKAWA, R. K. (1967). Random walk design in bioassay. *J. Amer. Stat. Assoc.* **62**, 824–856.

WETHERILL, G. B. (1963). Sequential estimation of quantal response curves. *J. R. Stat. Soc. Ser. B* **25**, 1–38.