NISS

Stratified Ordinal Regression: A Tool for Combining Information from Disparate Toxicological Studies

R.J. Carroll, D.G. Simpson and H. Zhou

Technical Report Number 26
September, 1994

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Although the information in this document has been funded wholly or in part by the United States Environmental Protection Agency under assistance agreement #CR819638-01-0 to the National Institute of Statistical Sciences, it may not necessarily reflect the views of the Agency and no official endorsement should be inferred.

STRATIFIED ORDINAL REGRESSION: A TOOL FOR COMBINING INFORMATION FROM DISPARATE TOXICOLOGICAL STUDIES

R. J. Carroll
Department of Statistics
Texas A&M University
College Station TX 77843–3143

D. G. Simpson
Department of Statistics
University of Illinois
Champaign IL 61820

H. Zhou
NISS
P. O. Box 14162
Research Triangle Park NC 27709

Abstract

This article focuses on risk assessment for acute inhalation exposure to perchloroethylene (PERC), and develops techniques for combining information from the available studies. The use of stratification is discussed as a way to address systematic aspects of data heterogeneity. It is demonstrated that failing to stratify on key variables can result in a misleading analysis. Humans appear to be much more sensitive to PERC than rats or mice given the same ambient air concentration. In fact, for the PERC data, the common practice of dividing the animal effective dose by ten seems close to the mark. An advantage of the analysis presented here is that it provides a way to scientifically investigate aspects of risk assessments such as species extrapolations. The same kinds of methods can be used for other chemicals for which sufficient data are available to estimate the effects of interest. [1]

[1]Research supported by the U.S. Environmental Protection Agency under Cooperative Agreement #CR819638-01-0, the National Cancer Institute under grant CA–57030, the National Science Foundation under contract DMS 92-07730 and the National Security Agency under contract MDA 904-92-H-3058. Daniel Guth of the USEPA provided the data and invaluable assistance in the preparation of this report.

1 INTRODUCTION

The Clean Air Act Amendments of 1990 require the Environmental Protection Agency to develop emission standards for 189 pollutants and to set standards for other substances "to provide an ample margin of safety to protect public health." Available data come from all relevant studies in the literature. Different studies may include different toxicological endpoints, and multiple endpoints appear within a given study. Moreover, the database includes multiple species, and there is tremendous variation in experimental protocol from one study to the next. The heterogeneity of the available data poses a serious challenge to the risk assessor.

Focusing on risk assessment for acute inhalation exposure to perchloroethylene (PERC), we develop techniques for combining information from the available studies. The cornerstones of the approach are: (1) the reduction of diverse endpoints to a common ordinal scale of severity categories; (2) partial stratification into more homogeneous subgroups of studies; and (3) uncertainty estimates that are adjusted for correlations between responses from the same study. The analysis provides tools for estimating risks in the subpopulations, and it allows the risk assessor to take advantage of the diversity of the data to examine species extrapolations and other sources of model uncertainty.

The use of ordinal severity scores is not unprecedented, see, e.g., Hertzberg and Miller (1985), Hertzberg (1989), and Guth, Jarabek, Wymer and Hertzberg (1991) who used ordinal logistic regression across studies, in some cases adjusting for species differences via an empirically derived "human equivalent concentration." A distinct advantage of adverse outcome modeling is that it provides a way to put very different quantitative measurements on a common scale. Hertzberg (1989) noted that, unlike the reference/benchmark dose methods, adverse outcome modeling provides risk estimates for doses other than the reference/benchmark dose. However, combining data from different studies requires careful consideration of study differences such as different species, endpoints, time frames, experimental purposes, etc. In particular, if one were to combine all the data in the PERC data base within a single species, and then do a categorical regression analysis, one would be assuming that, conditional on the exposure, the response scores are the result of a pure random (multinomial) sampling process. The assumption of a pure random (multinomial) error distribution across studies is difficult to defend scientifically or statistically.

The use of stratification in this context is a way to address systematic aspects of the heterogeneity of the data. In the design and analysis of experiments, stratification is an important and

powerful concept for eliminating bias, decreasing observed variation and hence improving both the accuracy and precision of inferences. We show that lack of proper stratification results in a misleading analysis which underestimates the sensitivity of humans to PERC.

The PERC data base consists of a number of different studies involving different investigators, species, genders, dose and duration of exposure; see section 2 for more details. Stratification as a conceptual tool attempts to control for these different factors by placing the different experimental outcomes into *strata*, or homogeneous subgroups. As a statistical procedure, stratification involves modeling which differentiates the different strata, while taking advantage of any resulting homogeneities in the subgroups.

Stratification is useful if there are identifiable *systematic* differences between studies. For example, different studies may use different test species, with allometric dose conversions to human equivalent concentrations (to account for species differences in respiration rates, etc.). Stratification allows for the possibility that there are systematic differences among species not explainable by a dose conversion.

There are other factors which may be candidates for stratification, although we do not pursue these in our analyses of PERC. Different endpoint categories might well lead to differences in dose-response, and stratification by endpoints may be useful to detect these differences. In addition, studies may have exposures to concentrations of different orders of magnitude, and stratification may be necessary to account for this.

In general, comparing a stratified model with an unstratified model provides a method for testing significance of the differences between strata, a so–called test of homogeneity. Moreover, differences among different subgroups may point to important differences in mechanism between different species, endpoint classes, or at different doses, whereas similarities between groups may provide support for extrapolations.

2 METHODS

2.1 Data acquisition

PERC is a widely used solvent, and its effects have been investigated in a number of small studies. We have collected the information from these studies and converted them into a PERC data base, the details of which will be reported elsewhere. In broad outline, initially a literature search was done to find all available data from published sources, proceedings and technical reports. The initial

set was screened to remove poorly documented studies. Since PERC is widely used, there exist human studies at low levels of exposure. Test species were mice, rats, rabbits, dogs and humans. The rabbit and dog groups were omitted from the analysis due to their scarcity in the data base.

A profile of these studies is given in Table 1. In this table, we distinguish between a *severity* study, which is clearly aimed at risk assessment for acute exposures, and a *mortality* study, which is aimed at fatal exposures and which has outcomes reported merely as survival or death. There were no pure mortality studies in the PERC data base, but a number of studies included lethal levels of exposure. In one of these studies a portion of the animal responses were reported simply as lethal or nonlethal, with no information on nonlethal adverse outcomes. In this case the nonlethal outcomes are censored.

Each study consisted of a number of groups of test subjects, with concentration and duration of exposure reported. The number of test subjects in each group varied, and we have reported the average number in each species-sex combination.

2.2 Ordinal Response Modeling

In the present analysis responses are measured at the group level via a categorical severity score. This reduces the various endpoints in different studies to ordinal severity categories: no adverse effect (N), mildly adverse outcome (A), and lethal or severe outcome (S). Thus, for example, a study might report the results on a group of six rats, and as our response we used a summary severity category for the whole group. Simpson, Carroll and Xie (1994) described latent variable models for group responses. They also considered the PERC data, and found that the latent individual responses within the dose groups were so highly correlated that for statistical purposes one could treat a group as a single individual. That is the approach taken here.

Severity judgements were on biological rather than statistical considerations. The use of statistical tests of significance at this stage would bias the subsequent ordinal regression analysis.

Concentration and duration of exposure are the primary independent regression variables for predicting the probabilities of the different severity categories. The primary covariates in the stratified analysis are species and gender. These were chosen as obvious candidates for illustration purposes. Further analysis may reveal other covariates of interest. See the appendix for details of the statistical methods.

2.3 Interval Censoring

Interval censoring occurs when the value of a measurement is known only to lie in an interval of values. If the response is ordinal, interval censoring occurs if certain observations are known only to be in one of several categories. The simplest example occurs if we combine a mortality study, in which the response is death versus survival, and a severity study with three categories corresponding to no adverse response, moderately adverse response and frank effect. Survivors in the mortality study are interval censored with respect to the three category scoring system, because the information on non–lethal adverse responses is missing. Another type of interval censoring occurs if biological thresholds between adjacent severity categories cannot be determined for a reported continuous response endpoint. We use the method of Carroll, Simpson & Zhou (1994) to handle interval censored responses in the categorical regression analysis.

2.4 Equivalent Doses (ED10's)

Traditional toxicological risk assessments derive a single acceptable dose from a selected study, typically on animals, and then make a series of extrapolations to obtain an acceptable dose for humans. One approach is to fit a model to the available data, and then compute an estimate of the ED_{10}, the concentration at a given duration which leads to a 10% adverse (AEL) or lethal (FRANK) effect rate. To give added protection for uncertainty, it is typical to compute a lower 95% confidence bound on the ED_{10}, and then divide this lower bound by a factor of 10 or 100, both steps attempting to account for uncertainty in the modeling procedure. This procedure outlined above tends to be conservative, not least of which because of the use of a single study usually causes the 95% lower bound for the ED_{10} to be smaller than if one were to make use of all available data. By combining data from a number of different studies, as well as accounting directly for modeling error, we hope to provide more accurate risk estimates. In this analysis, we used the ED_{10} as a measure of risk.

2.5 Statistical uncertainty estimates

Because of the correlations between responses from the same study, the usual variance estimates, which assume complete independence, are invalid. We adjust for correlation by the method of generalized estimating equations as described by Carroll, Simpson and Zhou (1994). This method provides more broadly valid confidence intervals than would be obtained by assuming independence.

The idea of the method is to replace the usual variance estimates and confidence intervals associated with maximum likelihood estimation by more general formulas. Thus, the ordinal regression model is assumed to be correct after averaging over random interstudy effects, but the analysis does not assume independence of different groups within a study. Typically the generalized analysis results in wider confidence intervals than a naive analysis that assumes independence. This reflects the fact that animals within a given study tend to respond more similarly to each other than to animals of the same species from other studies. Such differences are often described as interlaboratory effects.

3 RESULTS

In our first analysis, we pooled all the data and fit the model (1) in section 5.1. Figure 1 contains two important features; (a) a plot indicating the exposure concentration (in ambient air), duration and severity category for each group in the PERC data base; and (b) the estimated ED_{10} and its associated 95% confidence band when pooling all the data. The "censored" category in (a) always refers to interval censoring obtained by pooling the nonadverse and adverse categories. Part (b) shows the negative slope that one would expect; as duration increases, the estimated ED_{10} should decrease. Note that the confidence intervals obtained from pooled analysis are extremely wide, because the reference population includes three different species, as well as vastly different experimental protocols, different laboratories, etc.

We next performed a stratified analysis, stratified on the basis of species. Referring to the statistical methods in section 5.2, we fit in turn models (3) and (5). The former model allows for stratum effects but assumes that the effects of concentration and duration do not depend on the strata. The latter model allows for stratum–specific concentration effects. Model (3) provided a statistically significant improvement over model (1) ($p < 0.001$), while model (5) provided a statistically significant improvement over model (3) ($p < 0.04$). There was little evidence of an important stratum–specific duration effect.

The addition of stratification variables in the model was statistically significant, and the effects appear to be practically significant as well. Figure 2 shows concentrations and durations, as well as the species associated with each point on the graph. The ED_{10} lines for the different species are based upon the model (5), which has stratum–specific concentration effects. The immediate conclusion is that the ED_{10}'s are typically an order of magnitude smaller for humans than they are for rats, and similarly for mice at low durations of exposure. Note also that the rat and human lines

are very nearly parallel on the log–log scale, with the mice line being only somewhat nonparallel. The line for mice is estimated with much less precision than the human and rat lines, and the observed nonparallelism must not be overinterpreted. Because the duration parameter is shared by all species, this parallelism among the species is a reflection of the similarity of the concentration slope parameters across species. Thus, the major differences between rats and humans especially, and to a less extent the mice, appear to be explained by differences in uptake rather than differences in mechanism. In particular, the differences might be addressed by scaling up the concentrations, because the scale is reflected in the intercept rather than the slope if concentrations enter the model logarithmically. For mice, there may be a difference in mechanism, but the lack of precision in the line for mice makes such a conclusion tenuous at best.

In Figure 3, we combined the results of the pooled and stratified analyses. The vertical lines represent the pooled ED_{10} and its 95% confidence interval and the median duration of exposure observed in the PERC data base. The horizontal lines are the stratum–specific ED_{10}'s. There are major differences here. For example, the estimated ED_{10} for rats falls outside the pooled confidence interval. Note the previously mentioned fact that the ED_{10} for mice is poorly estimated relative to that for humans and rats.

We have run further analyses on these data, by allowing for species/gender stratum effects. While we observed some statistically significant effects in such analyses, in general the practical effects were not striking. To illustrate this, consider Figure 4, which is the analogue to Figure 3 when gender and species are used to form strata. One observes in Figure 4 some small gender effects, but these do not appear to be of major consequence.

4 CONCLUSIONS

In the PERC data base, the effect of stratification is striking. We have seen that the pooled analysis misses important effects, primarily that humans are much more sensitive to PERC than rats, at least based upon the empirical human equivalent concentration conversions. The ED_{10}'s are an order of magnitude smaller for humans than they are for rats and mice, the latter at low durations of exposure.

We concentrate here on rats and humans. The near parallelism of the lines suggests that the observed differences between the two species may be due to one of two factors: (a) humans have different rates of uptake than rats; or (b) that the human exposures are much lower than for rats

so that we are observing different effective severity categories in the different species. For either interpretation, the analysis suggests empirical human equivalent concentration adjustments for rats when exposed to PERC. In fact, for the PERC data, the common practice of dividing the animal effective dose by ten seems close to the mark.

A major advantage of the analysis presented here is that it provides a way to scientifically investigate aspects of risk assessments such as species extrapolations. The same kinds of methods can be used for other chemicals for which sufficient data are available to estimate the effects of interest.

REFERENCES

Agresti, A. (1984). *Analysis of Ordinal Categorical Data.* Wiley, New York.

Anderson, D.A. & Aitkin, M. (1986). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society*, Series B, 47, 203-210.

Carroll, R. J., Simpson, D. G. & Zhou, H. (1994). Interval censoring and marginal analysis in ordinal regression. Technical report, National Institute of Statistical Sciences.

Cox, D.R. & Hinkley, D.V. (1981). *Theoretical Statistics.* Academic Press, London.

Dourson, M.L., Hertzberg, R.C., Hartung, R. & Blackburn, K. (1985). Novel methods for the estimation of acceptable daily intake. *Toxicology and Industrial Health*, 1, 23-33.

Guth, D.J., Jarabek, A.M., Wymer, L. & Hertzberg, R.C. (1991). Evaluation of risk assessment methods for short–term inhalation exposure. Proceedings of 84th Annual Meeting of the Air & Waste Management Association.

Hertzberg, R.C. (1989). Fitting a model to categorical response data with application to species extrapolation of toxicity. *Health Physics*, 57, 405-409.

Hertzberg, R.C. & Miller, M. (1985). A statistical model for species extrapolation using categorical response data. *Toxicology and Industrial Health*, 1, 43-63.

Hertzberg, R.C. & Wymer, L. (1991). Modeling the severity of toxic effects. Proceedings of 84th Annual Meeting of the Air & Waste Management Association.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society,* Series B, 42, 109-142.

Simpson, D.G., Carroll, R.J. & Xie, M. (1994). Scaled link functions and quasilikelihood inferences for grouped binary data. Technical Report, National Institute of Statistical Sciences, Research Triangle Park, NC.

Zwart, A. & Woutersen, R.A. (1988). Acute inhalation toxicity of chlorine in rats and mice: time–concentration–mortality relationships and effects of respiration. *Journal of Hazardous*

Materials, 19, 195-208.

5 STATISTICAL APPENDIX

5.1 Homogeneous Logistic Regression

Assume S severity categories labeled 0, 1, ..., S, for example, three categories corresponding to no effect, mild adverse effect, and frank effect. Let Y be the ordinal severity category for a particular group, and let $x_1 = \log_{10}(\text{concentration})$ and $x_2 = \log_{10}(\text{duration})$ be the group's exposure history. Hertzberg & Wymer (1991) proposed to model Y using polytomous logistic regression,

$$\Pr(Y \geq s \mid x_1, x_2) = H(\alpha_s + \beta_1 x_1 + \beta_2 x_2), \quad s = 1, 2, \ldots, S, \tag{1}$$

where $H(u) = e^u/(1+e^u)$ and $H^{-1}(p) = \log\{p/(1-p)\}$, the so-called "logit" of p. If $S \geq 2$ then (1) entails a proportional odds assumption for the different severity categories; see McCullagh (1980) and Agresti (1984).

Model (1) yields a formula for the 100q% effective dose for severity category s, i.e., the dose level with 100q% risk of a response as severe as category s. On the logarithmic scale the effective dose is

$$\text{ED}_{100q}(x_2) = \frac{\text{logit}(q) - \alpha_s - \beta_2 x_2}{\beta_1}. \tag{2}$$

For a given duration, lower values of ED_{100q} correspond to more potently toxic chemicals. If the chemical is toxic a plot of ED_{100q} versus the log–duration yields a downward sloping line; smaller doses can be effective if the duration of exposure is increased. Inserting estimates for the unknown parameters yields an estimate of the ED_{100q} line. Often the delta method (Cox & Hinkley, 1981) is used to get pointwise confidence intervals for the ED_{100q}; our analysis used this procedure.

5.2 Covariates and Stratification

To investigate systematic sources of heterogeneity in the data base, we allow certain parameters to vary between different subsets of the data. These subsets are called *strata*, and they are constructed with the hope that the data they contain are more homogeneous than the data base as a whole. In addition, we can incorporate covariate information if available, to account for potential bias in the model.

We therefore expand model (1) to include a vector z containing stratification variables and covariate information. The vector z might include mechanistic information, but it will nearly always include indicator variables for strata corresponding to different species, sexes, etc. The expanded model is given by

$$\Pr(Y \geq s \mid x_1, x_2, z) = H(\alpha_s + \beta_1 x_1 + \beta_2 x_2 + \gamma' z), \quad s = 1, 2, \ldots, S. \tag{3}$$

Here $z = (z_1, z_2, ...)'$ and $\gamma = (\gamma_1, \gamma_2, ...)'$ Assuming that the stratification variables z are approximately independent of dose and duration (after adjustment of the former to human equivalent concentrations), the formula for the effective dose given the settings of all the other variables is

$$\mathrm{ED}_{100q}(x_2, z) = \frac{\mathrm{logit}(q) - \alpha_s - \beta_2 x_2 - \gamma' z}{\beta_1}. \tag{4}$$

The model (3) can be further expanded, for example to allow different strata to have different log–concentration parameters. For example, if there are $j = 1, ..., J$ different strata, then $z = (z_1, ..., z_J)$, where $z_j = 1$ if an observation comes from stratum j and $z_j = 0$ otherwise. The model is

$$\Pr(Y \geq s \mid x_1, x_2, z_j = 1) \;=\; H(\alpha_s + \beta_{1j} x_1 + \beta_2 x_2 + \gamma_j), \quad s = 1, 2, \ldots, S,\ j = 1, 2 \ldots, J, \tag{5}$$

where it is understood that $\gamma_J = 0$, a convention necessary to make the model statistically identifiable. For such a model, the effective dose for stratum j is

$$\mathrm{ED}_{100q}(x_2, z_j = 1) = \frac{\mathrm{logit}(q) - \alpha_s - \beta_2 x_2 - \gamma_j}{\beta_{1j}}. \tag{6}$$

Models (3) and (5) can be fit using standard statistical software, for instance, using the SAS Logistic procedure (SAS Institute Inc.). Its structure is very general, accommodating many different kinds of partial stratification and covariate information. The possibilities include stratification of the baseline risk on species, sex, endpoint category, etc.; stratification of the potencies reflected in the β parameter, e.g., different sensitivities for different types of endpoints; and empirically estimated cross–species dose conversions. With additional computational effort $\gamma' z$ may be replaced by a nonlinear function $f(\gamma, z)$, e.g., mechanistic information. One way to investigate and reduce modeling error is to fit a hierarchy of models like (3) and (5) incorporating increasingly fine stratification and do model selection within the series. Model selection within this framework is commonly directed by a series of likelihood ratio tests between nested models. It should be noted that multiple testing between a number of models inflates the false positive rate, so it is possible to get nominally statistically significant differences that are non–significant after adjustment for multiple testing. Even results that are highly significant in the statistical sense can be relatively insignificant in the scientific sense. The graphical techniques used in the examples can reveal these kinds of phenomena.

	Severity Studies			Mortality Studies		
	# of Studies	Total # of Groups	Average # per Group	# of Studies	Total # of Groups	Average # Per Group
Mice - F	5	47	5.87	1	8	8
Mice - M	3	20	14.85	0	0	0
Mice - B	2	12	13.33	0	0	0
Rats - F	4	20	5.35	0	0	0
Rats - M	2	14	11.43	0	0	0
Rats - B	1	19	16.12	0	0	0
Human - M	3	18	5.67	0	0	0
Human - B	2	7	6.29	0	0	0

Table 1: *Information on Perchloroethylene (PERC) Here "–F" means females, "–M" means males and "–B" indicates that gender was either unspecified or the group was mixed.*

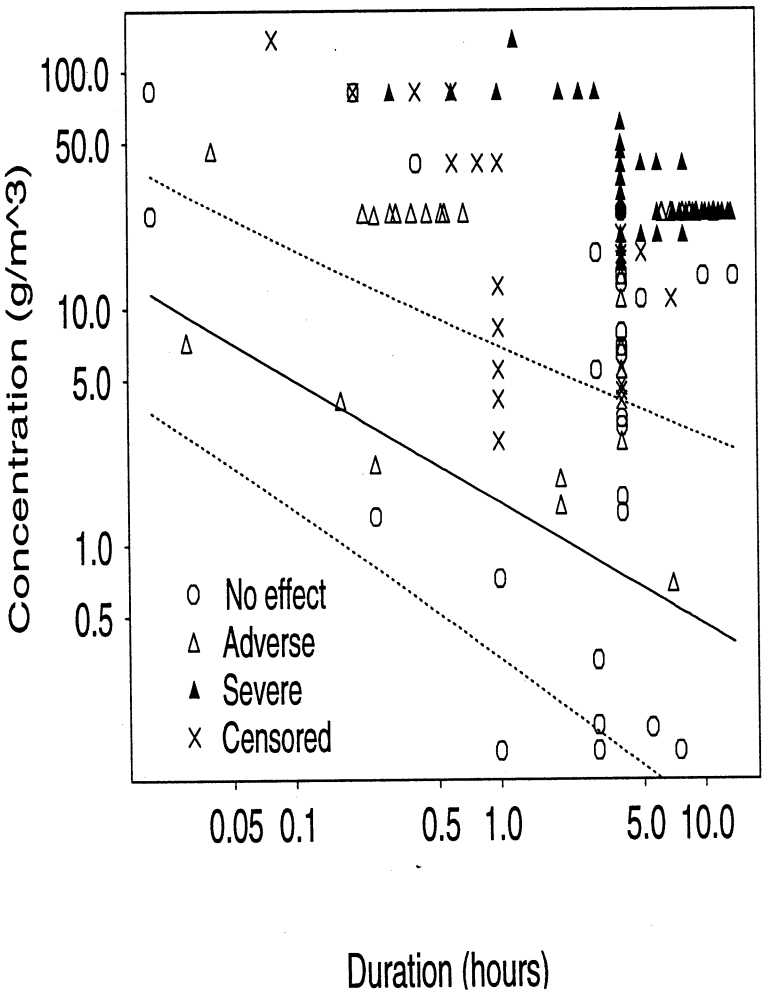

Figure 1: *PERC data, ED10 line (solid line) when pooling all studies, with associated 95% confidence bands (dashed lines).*

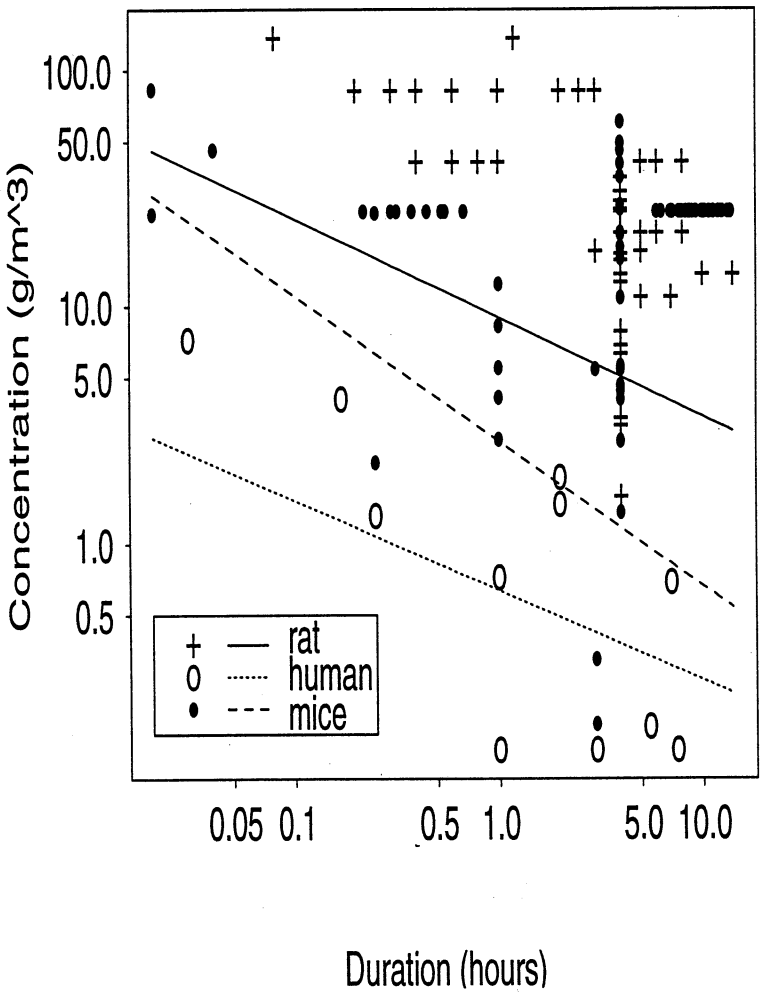

Figure 2: *PERC Data, ED10 lines when intercepts and slopes are stratified by species.*

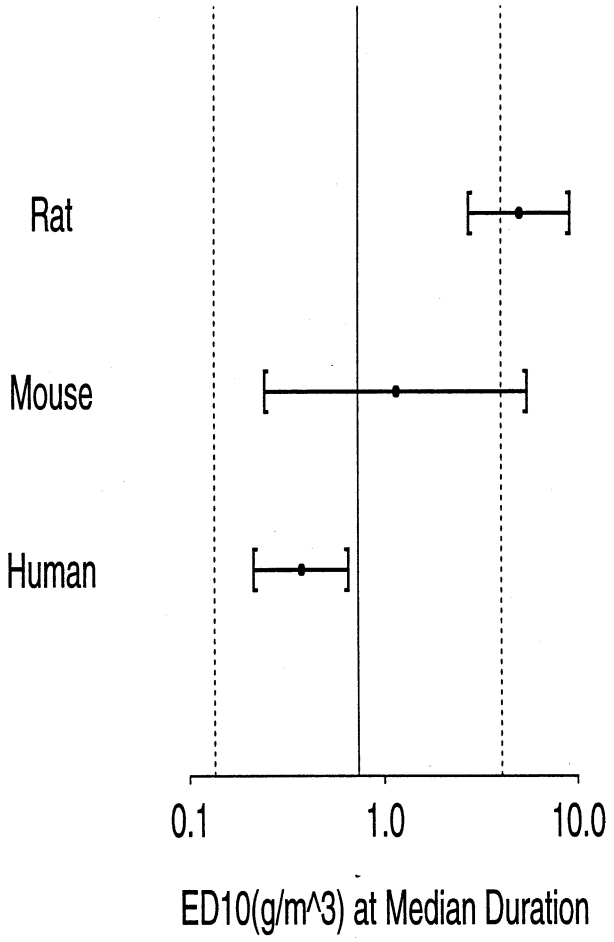

Figure 3: *PERC data, log (base 10) ED10's for different species combinations at the median duration of all studies. The solid verticle line is the log (base 10) ED10 when all studies are pooled, while the dashed vertical lines are the associated 95% confidence intervals..*

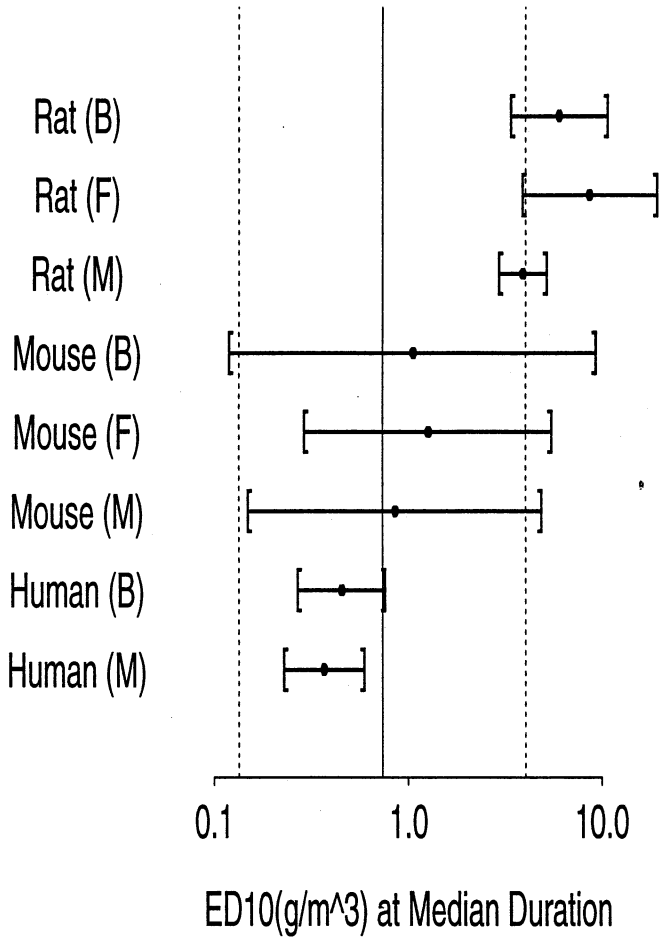

Figure 4: *PERC data, log (base 10) ED10's for different sex and species combinations at the median duration of all studies. The solid verticle line is the log (base 10) ED10 when all studies are pooled, while the dashed vertical lines are the associated 95% confidence intervals. Here "–f" means females, "–m" means males and "–b" indicates that gender was either unspecified or the group was mixed.*