



5175 NE River Rd
Sauk Rapids, MN 56379
Tel: (320) 281-0676
Fax: (320) 323-4418
jkern@KernStat.com

To: Mr. Allan Taylor and Dr. Deborah Mackenzie-Taylor
Michigan Department of Environmental Quality
Waste and Hazardous Materials Division

From: John W. Kern Ph.D.

Re: Review of University of Michigan Dioxin Exposure Study Statistical Modeling Methods, Analysis, and Interpretation.

CC:

Date: 3/3/2009

1.0 Introduction

The University of Michigan Dioxin Exposure Study (UMDES) is a large observational study intended to identify factors associated with levels of dioxin like compounds (DLCs) in human blood serum in the Midland-Saginaw area of Michigan. The Dow Chemical Company (Dow) and others have suggested that these studies provide results that could be useful to the Michigan Department of Environmental Quality (MDEQ), the Michigan Department of Community Health (MDCH), and Dow in managing risks of DLC exposure in the Midland and Saginaw areas where DLCs are known to persist in soil, sediments, and biota. The MDEQ has requested Kern Statistical Services, Inc. (KERN) to review and evaluate the UMDES for this potential application with focus on:

- assessing the utility and applicability of the UMDES results;
- identifying how and in what context(s) they may be used; and if appropriate,
- determining if modifications are necessary to improve their utility in risk management decisions.

By including discussions of risk management herein, KERN does not imply that the MDEQ will use UMDES findings in regulatory decision making about Dow's correction action.

2.0 Summary of Findings

The UMDES represents one of the largest efforts to study DLC concentrations in human serum; however, until further evaluation has been completed, the results of the UMDES should not be used or relied on for decision making. Following is a summary of issues that impact the interpretability of the UMDES study results:

- Data collected as part of the UMDES are not publicly available beyond the UMDES research team due to confidentiality requirements. This limitation on data accessibility makes it difficult for the

MDEQ and MDCH to evaluate the completeness and utility of these data and the UMDES research results for making risk management decisions in the Midland-Saginaw areas.

- The types of statistical methods used in the UMDES are not appropriate for use with the UMDES design. The methods used in the UMDES are appropriate for application in controlled experiments, in which a set of *a priori* models have been clearly identified. The UMDES is not a controlled experiment and would more appropriately be classified somewhere between exploratory and confirmatory observational studies because *a priori* models were not established and extensive data reduction with automated variable selection was conducted. This apparent mismatch between the type of study design and selection of statistical methods has resulted in misapplication of statistical methods for the purpose of variable selection and data reduction. The problems with this type of approach are well documented in the statistical literature (Altman and Andersen, 1989; Derksen and Keselman, 1992; Freeman, 1999; Grambsch and O'Brien, 1991; Harrell, 2001) and directly limit the utility of the UMDES results.
- The sampling design and selection of research subjects may not adequately represent critical target populations (i.e., those with the highest exposures to DLC contamination from Dow). Consequently, any statistical inferences with regard to these critical target populations should be made with caution as they may be highly uncertain. This point was raised by the MDEQ in 2004 (MDEQ, 2004) and at various subsequent points in discussions prior to finalization of the sampling design.
- The statistical modeling approach is complex and consists largely of automated variable selection among large numbers of potentially important explanatory variables. The flaws of automated variable selection procedures are widely known to inhibit both model interpretability and predictive capability (Altman and Andersen, 1989; Derksen and Keselman, 1992; Freeman, 1999; Grambsch and O'Brien, 1991; Harrell, 2001).
- The statistical methods and study design are not appropriate for ranking the importance of environmental sources for explaining serum DLC levels.
- Applicability of the UMDES results could be improved if associations between serum DLC and each environmental source were described individually with estimates of the strength of the relationship as well as measures of uncertainty.

Each of the above major issues is sufficient to warrant further evaluation of the UMDES results before any interpretation occurs and/or results are applied for risk management purposes.

3.0 Applicability of Research to Risk Management

Scientific research conducted according to “the scientific method”, is an iterative process starting with *a priori* formulation of research questions, followed by study design and sample selection, careful and detailed statistical analyses, and formulation of new research questions and insights based on these data and analyses. In this process, findings are developed and confirmed or rejected through subsequent iterations. Spurious results (false negative and false positive errors) are identified when results are not repeatable. When results are repeatable, prior beliefs are strengthened. In this general context, false positive errors are defined as those that would falsely identify significant effects; false negative results would fail to identify significant effects.

In contrast to scientific research, risk management is a process of integration of diverse sources of information for selection among remedial alternatives that, unlike academic research findings, are often not reversible. Remedial actions based on false positive research results may be needlessly expensive and may be accompanied by unintended negative consequences. Failure to take action due to false negative results may result in failure to adequately protect public health and the environment. This distinction between research and risk management influences how users of the UMDES must interpret study results. Risk managers have fewer iterative cycles with which to refine research questions and to answer them, and false positive (negative) interpretations have costly and, at times, immediate consequences.

Due to the above concerns, it is necessary that study results are fully transparent to risk managers and well tested and validated prior to application. Research findings that have not been fully validated through the iterative research cycle must be applied with a healthy dose of precaution. This need to take a precautionary stance with new research, combined with the lack of public availability of data, makes it particularly difficult to apply UMDES results to risk management decisions at hand in the Midland-Saginaw area.

In spite of these limitations, it is anticipated that with certain additional analysis efforts, results of the UMDES can add significantly to the understanding of factors associated with blood serum DLC levels in the Midland-Saginaw areas. Based on review of materials publicly available at the UMDES web site (UMDES, 2009), it appears that an appreciable quantity of data exist, some of which may be of sufficient quality to support evaluation of some lines of evidence associated with DLC exposure pathways for residents of Midland-Saginaw. The reports that currently reside on the UMDES web site suggest that results may not be directly interpretable without substantial clarification of the inner workings of the supporting data and statistical modeling efforts. The following section describes in more detail the issues that currently limit applicability of the UMDES studies to risk management.

4.0 Study Design

Study designs should incorporate the following components: 1) careful specification of objectives and research question(s); 2) selection of the type of study design that can best be implemented (e.g., controlled or observational studies); 3) development of a sampling design that is consistent with the research question(s) and intended statistical methods; and 4) identification of statistical methods that are consistent with each of the above. Each of these items is described below in more detail.

4.1 Specification of the Research Question(s)

There are a wide range of research questions described in the UMDES materials with varying degrees of relevance to risk management in the Midland-Saginaw areas. There are three primary goals stated in the UMDES materials that appear to be of greatest relevance to MDEQ risk management decisions.

Statement 1. *The study is being undertaken in response to concerns among the population of Midland and Saginaw Counties that dioxin related compounds from the Dow Chemical Company facilities in Midland have resulted in contamination of the City of Midland and have contaminated sediments in the Tittabawassee River Floodplain. There is concern that people's body burdens of dioxins, furans and PCBs are elevated because of environmental contamination.*

This relatively general statement suggests that the study would be designed to estimate associations between body burdens and environmental contaminant distributions in the Midland-Saginaw area. This is a reasonable general statement of the primary study objective, but ultimately reported results were broader than the study design and statistical analyses could support. The following two additional statements of study objectives illustrate this.

Statement 2. *A central goal of the study is to determine which factors explain variation in serum congener levels, and to quantify how much variation each factor explains.*

This goal to partition serum DLC variation into variance components is in general not achievable with an observational study design. Because associations between explanatory variables (discrete and continuous) cannot be controlled, as in a designed experiment, the proportion of variation explained by each variable is conditional on variables included or excluded from statistical models. The UMDES has used these invalid variance partitioning results to support the unequivocal assertion that "Soil and dust were not important contributors to serum TEQ, PCDDs, PCDFs, or PCBs."

Statement 3. *...to find out whether the elevated levels of dioxins in the soil in the city of Midland, and in the Tittabawassee River flood plain between Midland and Saginaw, have also caused elevated levels of dioxins in residents' bodies.*

Again, because the UMDES studies are observational, cause and effect cannot be inferred. Associations can be described, but causation, or lack thereof cannot be established from observational studies.

In general, these study goals are not clear statements of research hypotheses. This failure to clearly state research hypotheses has led to large data reduction and model selection calculations in efforts to develop coherent descriptions of previously unanticipated data relationships. In contrast, it would have been expected that critical explanatory variables would have been identified for evaluation of specific objectives. Rather, statistical modeling was conducted to “sift” through the multitude of explanatory variables with little regard for the above-stated objectives.

It is recommended that the UM study team simplify their statistical approach by limiting and/or prioritizing hypotheses to be tested. A relative priority should be established for each explanatory variable, with critical explanatory variables (i.e., those necessary to evaluate the above-stated objectives related to DLC contamination from Dow) given the highest priority for statistical analyses. More detailed research questions should be specified through identification of appropriate null and research hypotheses for each critical explanatory variable.

The results of not clearly specifying objectives and research questions at the beginning of a study are numerous: the study generally is inefficient, too many or too few factors are often tested, important or critical factors are overlooked while other less important factors are evaluated, and substantial errors and uninterpretable results are generated due to multicollinearity and other problems with model selection. Many of these problems have been identified in the UMDES, as described in the following sections.

4.2 Observational Study Design

Appropriate statistical analysis and the scope of inference are largely determined by study design. Neter et al (1996) classifies study designs into four groups: controlled experiments, controlled experiments with supplemental variables, confirmatory observational studies, and exploratory observational studies. Exploratory observational studies are typically performed when controlled experiments cannot be conducted, theoretical models describing relationships are not available, and important variables may be unobservable. In these situations a large number of variables are often measured in hopes of identifying important relationships between one or more of these variables and the dependent variable of interest (i.e., the response variable). In contrast, confirmatory observational studies are usually focused on a set of “primary variables” that have been identified previously as important predictors of the response. Additional variables thought to be of some importance are also included to aid in explaining variance in the response variable with the hope of increasing precision of effects estimates for primary variables. Data reduction is typically not part of a confirmatory observational study because

hypotheses and secondary variables have been identified prior to conducting the study. Confirmatory observational studies are often conducted as a follow up to test hypotheses formed through an exploratory observational study. Carefully controlled confirmatory studies are needed to infer cause and effect.

The UMDES studies are classified somewhere in between the exploratory observational study and the confirmatory observational study. *A priori* hypotheses have been identified for some variables (e.g., those known through the literature to influence blood serum levels such as age and body mass index or BMI); however, extensive variable reduction (i.e., through automated regression modeling procedures) was also used in the analysis to sort through the many combinations of possible explanatory variables. This “sifting” of the large number of candidate variables indicates a lack of a well formed set of *a priori* models, tilting the study more toward an exploratory study. In this case, the extensive literature on DLC exposure pathways could have been more fully integrated into development of a much smaller number of plausible models with reduced sets of explanatory variables to evaluate. Furthermore, as described in Section 4.1, it is necessary to identify critical explanatory variables (i.e., those that represent DLC contamination from Dow) and establish corresponding hypotheses to be tested by these models.

For example, it’s expected that serum DLC would be associated with physical condition of individual subjects, so variables such as serum lipid content, age and BMI would be included in regression models in order to improve precision of estimation. This would be termed “controlling” for age, lipid and BMI. In addition to these variables one would also include a critical variable of interest, such as DLC concentration in soil (C_{soil}).

The model would be of the form

$$\ln(C_{serum}) = \beta_0 + \beta_1 \ln(Serum\ Lipid) + \beta_2 \ln(Age) + \beta_3 \ln(BMI) + \beta_4 \ln(C_{soil}), \quad (1)$$

where natural logs are used to account for the skewness that is characteristic of environmental samples.

Evidence of association between serum and soil DLC is provided by the magnitude and precision of the estimated coefficient β_4 which can be summarized by a confidence interval. This information would provide risk managers with an understanding of the likely strength and uncertainty of estimated associations between serum and soil DLC levels.

A similar model could be considered for estimating the association between serum DLC levels and consumption of Tittabawassee River fish

$$\ln(C_{serum}) = \beta_0 + \beta_1 \ln(Serum\ Lipid) + \beta_2 \ln(Age) + \beta_3 \ln(BMI) + \beta_4 \ln(I_{fish}) \quad (2)$$

where I_{fish} indicates those survey respondents that consume Tittabawassee River fish. Evidence of association between soil DLC levels or fish consumption would again be judged based on the magnitude of the estimated coefficient and uncertainty of the estimate for β_4 . These models clearly separate “nuisance” variables from critical variables related to research questions and they are handled differently in the statistical analysis and reporting of results.

4.3 Selection of Subjects

In the UMDES studies, selection of research subjects, particularly in the Floodplain population defined by Garabrant et al (2005), was not adequate to represent the critical target populations defined by MDEQ (2004). As described by MDEQ (2004), critical target populations are those “most likely to have the highest exposures to DLC contamination from Dow.”

The Floodplain population was defined to include people who live on or near the 100-year floodplain of the Tittabawassee River (Garabrant et al, 2005). However, the population most likely to have elevated body burdens of DLCs due to elevated soil concentration is the subset of people who live on frequently-flooded portions of the Tittabawassee River floodplain or those areas of the floodplain that flood at least every seven to ten years or more frequently. As documented by MDEQ (2004), the highest soil concentrations are consistently located in these frequently-flooded portions of the 100-year floodplain. The UMDES Floodplain population includes a large number of people who do not live in the more highly-contaminated, frequently-flooded areas of the 100 year-flood plain. The UMDES Floodplain population, therefore, would be expected to dilute estimated effects due to exposure to DLC in soil. Consequently, effects within critical target populations are likely underestimated.

A way to address this problem would be to select participants in all study groups based on soil concentrations of DLCs. Instead, a “two-stage area probability household sample design” was used to select participants from the five geographically-designed study populations, including the Floodplain population, without regard to soil concentration. Inclusion of insufficient numbers of participants with high or low DLC concentrations would decrease the power to detect a significant relationship between soil and serum DLC concentrations.

Furthermore, other study subjects (e.g., high end fish consumers, game consumers, and other animal products consumers associated with the Tittabawassee River, Saginaw River, or Saginaw Bay) are not adequately represented. These critical food chain exposure factors are not necessarily related to the geographically-based study groups identified in the UMDES. Selection of participants in all study groups should be conditional on exposures to DLC contamination from Dow. Identification and inclusion of sufficient numbers of participants with exposures to both higher and lower DLC levels is essential to detect relationships between exposure to DLC contamination from Dow and serum concentrations with a reasonable level of statistical power.

4.4 Statistical Methods: Model Selection

The majority of regression analyses reported at the UMDES web site are based on multiple variable generalized linear models (McCullagh and Nelder, 1999), with a heavy reliance on stepwise model selection methods based on significance testing. Generalized linear models are well established in the applied and statistical literature, forming the primary tool-set for most modern statisticians, but when used in combination with automated variable selection procedures can be unreliable. Although widely available through reputable software implementations, the flaws of automated variable selection procedures are widely known. Harrell (2001), Grambsch and O'Brien (1991), Altman and Andersen (1989), Derksen and Keselman (1992), and others have studied the performance of various automated variable selection procedures and found a variety of pathologies inhibiting both model interpretability and predictive capability.

Model selection through the use of automated stepwise computer procedures based on significance testing (i.e., variable selection by p-value) is widely understood to result in models that are too complex, have poor out-of-sample predictive value, and often result in biased model coefficients that are uninterpretable and may be misleading. Freeman (1999) summarizes the implications of automated variable selection procedures, stating

"If causation or biological interpretability is the reason for fitting a multivariate model, then automatic stepwise algorithms will not provide interpretable answers, may produce substantial errors, and should not be employed."

Dissenting opinions toward Freeman's position are difficult to find in the statistical literature.

The primary objective of the UMDES is to develop causative or at least strong associations between environmental factors and blood serum levels of DLCs, if they are present. For example, Garabrant (2008) draws the following conclusion regarding the association between soil and blood serum DLC levels:

"Soil and dust were not important contributors to serum TEQ, PCDDs, PCDFs, or PCBs."

This result is stated unconditionally, in spite of the fact that it is conditional on the sampling design and the other variables included in, and excluded from, the final model selected through automated model selection procedures. These automated selection methods are known to perform poorly, particularly when interpretation of results is of central importance. While this result may be completely accurate, it is important to know if the soil relationship is sensitive to the other variables included in, or excluded from, the model. To evaluate the reliability of this result, it is necessary to fully evaluate the presence of multicollinearity in the explanatory variables and to specify other plausible candidate models for comparison. Other combinations of variables may fit the data equally well, yet result in different conclusions with regard to the importance of soil DLC concentrations.

An example of probable multicollinearity can be seen in the use of multiple fish consumption variables in the UMDES statistical models. Since these “independent variables” are likely correlated with each other, model results are flawed and uninterpretable. For example, in at least one reported model, fish consumption in general was associated with elevated serum DLC, while consumption of Tittabawasee fish was associated with lower serum DLC.

Another example is provided by the association between DLC concentrations in serum and soil, and the contradictory conclusions that were generated using more than one automated variable selection procedure (i.e., forward stepwise and backward elimination methods). Furthermore, at the November 2008 meeting of the Science Advisory Board, specific examples were identified in presentations that illustrated the correlation between soil DLC concentration and region of residence. These are just a few examples that strongly suggest the presence of multicollinearity in the UMDES statistical models. Based on the nature of other explanatory variables in the models that have been presented, it would be expected that additional redundancies may exist that may negatively impact automated model selection procedures and subsequent inferences based on final models.

The underlying problems that tend to cause automated model selection procedures to perform poorly stem from their deviation from the basic assumption underlying multiple variable regression analyses— independence of explanatory variables. Violation of this assumption limits the reliability of inferences from multivariable models when one or more explanatory variables are associated with explanatory variables of primary interest (i.e., those that measure contamination due to Dow) as well as the dependent variable, even when variable reduction is not employed.

In recognition of the importance of this issue and others that complicate the analysis of observational studies, the United States Environmental Protection Agency convened a panel to discuss these issues (Bateson et al, 2007) specific to epidemiologic analyses of air quality. Concerns expressed by this panel were consistent with this review of the UMDES. Confidence in the reliability of the UMDES conclusions can be developed through careful investigation including detailed interactions between the UMDES team and technical experts from the MDEQ and MDCH. A clear understanding of the stability, reliability, strengths, and limitations of the UMDES results is essential to their application to risk management.

4.5 Alternative to Variable Selection and Hypothesis Testing

An alternative to variable selection procedures is to first recognize that inclusion of correlated variables into a single regression model will result in unreliable results. Second, the importance of the two exposure pathways should be evaluated based on direct comparison of models, rather than by significance testing within a single regression model. Any number of models can be compared directly using information theoretic statistics such as the Akaike’s Information Criterion or AIC (Akaike, 1974). This approach provides a means to rank the relative strength of models leading to an understanding of their predictive power and to describe the uncertainty in conclusions that might inform risk management.

Importantly, this approach is not susceptible to the instabilities of variable selection procedures. Additionally, groups of models that have similar predictive power (i.e., similar AIC values) can be identified to risk managers allowing uncertainty in study results to inform decision makers. To date, the UMDES results have reported a single individual “winning” model without any indication of how other candidate models may fit the data. This is important information to risk managers because models that fit the data approximately as well as the single winner, but with different explanatory variables, may lead to very different risk management strategies.

5.0 Recommendations

Each of the shortcomings described in Section 4.0 limit the utility of the UMDES results. Each of these shortcomings will need to be addressed before the UMDES could potentially be used in support of risk management decisions:

- Specification of key research questions and identification of critical explanatory variables.
- Documentation of limitations in answering critical research questions due to sampling design and selection of subjects.
- Identification of appropriate statistical methods based on the key research questions, the type of study design being implemented (i.e., observational), and the available data.

Prior to development of statistical models, explanatory variables should be evaluated to describe the correlation structure among all explanatory variables. This approach will help to avoid issues of multicollinearity and confounding.

With respect to the potential for confounding, a set of plausible “candidate” models that are designed to test hypotheses for critical explanatory variables should be estimated and ranked based on model fit using AIC or other appropriate information theoretic statistics. Furthermore, specific analyses should be designed to be applicable to supporting lines of evidence related to risk management decisions in areas of DLC contamination associated with Dow in the Midland-Saginaw area.

Statistical methods for developing and evaluating a set of candidate models are described in Burnham and Anderson (1998). Central to this approach is abandoning the use of automated variable selection methods in favor of careful *a priori* specification of candidate models (hypotheses) of particular interest. These candidate models can be ranked based on information theoretic criteria such as the AIC (Akaike, 1974). Uncertainties in model interpretation are explicitly described and made transparent to the risk manager. Harrell (2001) also provides modeling strategies for selection of models that balance complexity with model fit, including diagnostic measures that are only recently becoming known to practitioners.

Development of inferences in this way requires careful thought and engagement between discipline experts and statisticians. To enhance the usefulness of the UMDES, it is necessary that the agencies have the opportunity to work with the UMDES team at the technical level to specify critical explanatory

variables, research questions, and candidate statistical models necessary to address the stated goals of the UMDES and to provide information critical to risk management decisions. It is fully expected that such a collaborative process would be useful in extending the utility of the UMDES studies from research to more directly informing risk management decisions in the Midland-Saginaw areas.

6.0 References

- Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automated Control AC* **19**, 716-723.
- Altman, D.G. and P.K. Andersen, 1989. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine*, **8**: 771-783.
- Bateson, T.F., Coull, B.A., Hubbell, B., Ito, K., Jerrett, M., Lumley, T., Thomas, D., Vedal, S., and M. Ross. 2007. Panel discussion review: Session three—issues involved in interpretation of epidemiologic analyses—statistical modeling. *Journal of Exposure Science and Environmental Epidemiology*. **17**, S90-S96.
- Burnham, K.P. and D.R. Anderson, 1998. *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer Verlag.
- Derksen, S. and H.J. Keselman. 1992. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, **45**: 265-282.
- Freeman, J. 1999 (2001 in text). Modern quantitative epidemiology in the hospital. In: Mayhall CG ed. *Hospital epidemiology and infection control*, 2e.. Philadelphia: Lippincott Williams & Wilkins, pp. 15-48.
- Garabrant, D. H., Franzblau, A., Gillespie, B., Lin, X., Lepkowski, J., Adriaens, P., and A. Demond. 2005. The University of Michigan Dioxin Exposure Study – Study Protocol. <http://www.sph.umich.edu/dioxin/Protocol/UMDES%20Overview%2003-06-05.pdf>
Last accessed December 12, 2008.
- Garabrant, D.H. 2008. *Project overview and results of linear regression models of serum dioxin levels*. Presented at Dioxin 2008, Birmingham, England. Last accessed December 3, 2008.
- Grambsch, P.M. and P.C. O'Brien. 1991. The effects of transformations and preliminary tests for non-linearity in regression. *Statistics in Medicine*, **10**:697-709.
- Harrell, Jr., F. E. 2001. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer-Verlag: New York.

McCullagh, P. and J.A. Nelder. 1999. *Generalized Linear Models, Second Edition. Monographs on Statistics and Applied Probability 37.* Chapman and Hall/CRC, New York.

MDEQ. 2004. Communication from Jim Sygo, Deputy Director of the MDEQ, to David H. Garabrant, Primary Investigator for the UMDES, September 28, 2004.

Neter, J., Kutner, M.H., Nachtsheim, C.J. and W. Wasserman. 1996. *Applied Linear Statistical Models, Fourth Edition.* Irwin Press, Chicago.

UMDES. 2009. University of Michigan Dioxin Exposure Study homepage:
<http://www.sph.umich.edu/dioxin/>. Last accessed February 15, 2009.