Pure

Syddansk Universitet

Statistical methods for the evaluation of health effects of prenatal mercury exposure

Budtz-Jørgensen, Esben; Keiding, N.; Grandjean, Philippe; Weihe, Pál; White, R.F.

Statistical Methods for the Evaluation of Health Effects of Prenatal Mercury Exposure

Esben Budtz-Jørgensen[1,2], Niels Keiding[1], Philippe Grandjean[2,4], Pal Weihe[2,3] and Roberta F. White[2,4]

[1]*Department of Biostatistics, University of Copenhagen*
Blegdamsvej 3, DK-2200 Copenhagen N, Denmark.
[2]*Institute of Public Health, University of Southern Denmark*
Winslowparken 17, DK-5000 Odense C, Denmark.
[3]*Faroese Hospital System, FR-100 Tórshavn, Faroe Islands*
[4]*Departments of Environmental Health and Neurology,*
Boston University Schools of Medicine and Public Health, Boston, MA 02118, USA

SUMMARY. Environmental risk assessment based on epidemiological data puts stringent demands on the statistical procedures. *First*, convincing evidence has to be established that there is at all a risk. In practice this endeavor requires prudent use of the observational epidemiological information with delicate balancing between utilizing the information optimally but not over-interpreting it. If a case for an environmental risk has been made, the *second* challenge is to provide useful input that regulatory authorities can use to set standards. This paper surveys some of these issues in the concrete case of neurobehavioral effects in Faroese children prenatally exposed to methylmercury. A selection of modern, appropriate methods has been applied in the analysis of this material that may be considered typical of environmental epidemiology today. In particular we emphasize the potential of *structural equation models* for improving standard multiple regression analysis of complex environmental epidemiology data.

KEY WORDS: Environmental epidemiology, Confounding, Measurement error, Multiple endpoints, Structural equation.

1 Introduction

Human health is affected by both inherited and environmental factors, and public-health experts have been particularly eager to identify specific hazardous environments that could be targeted

by preventative efforts. Statistical methods were not necessary when early observations suggested that cholera was caused by polluted drinking water, and many other early observations of environmental risks were made without any detailed statistical insight. With chemical exposures, early observations on poisoned victims likewise did not require sophisticated computations to demonstrate associations. As perhaps best illustrated by the asbestos experience, modern biostatistical methods subsequently allowed identification of delayed effects that had not been apparent before. Because of the attention to environmental factors, it seems clear that all the easy victories were already won long ago, with or without the help of biostatistics. Public health research now faces the serious challenge of identifying specific risk factors that may result only in non-specific, delayed effects where perhaps even individual predisposition plays a role. In dealing with this challenge, public health research must use the most sensitive and accurate methods to determine exposure levels, outcomes, and important confounders. Expanded biostatistics methodologies are needed to analyze the complicated interrelationships and to help identify the hazardous environmental factors.

In this paper, we illustrate the current challenges using environmental exposure to methylmercury as an example. Poisoning incidents have amply documented that this chemical accumulates in fish and seafood, and that it can cause serious damage to the nervous system, especially when extreme exposure occurs prenatally, i.e., due to the mother's diet during pregnancy. Although this hazard was therefore recognized, exposure limits were based on extrapolations from human poisoning incidents or experimental animal studies (WHO, 1990). New data was collected on subjects who had been exposed to prevalent environmental levels of methylmercury, as recently reviewed (NAS, 2000). The question therefore emerged how to extract the best possible information from these results using modern biostatistical methods.

This paper focuses on statistical methods which are relevant for deciding whether or not the exposure in question has an adverse effect. The further challenge concerning estimation of a safe exposure level of a substance suspected to be toxic was considered in Budtz-Jørgensen et al. (2001).

2   The Faroese Mercury Study

In the Faroe Islands, the population is exposed to increased levels methyl mercury mainly though consumption of contaminated pilot whale meat. During 1986-1987 a birth cohort of 1022

Faroese children was thefore established, and is being studied prospectively to examine possible adverse effects of prenatal exposure to methylmercury. The intrauterine methylmercury exposure was determined by analysis of umbilical cord blood and maternal hair for mercury (Grandjean et al., 1992). Furthermore, a midwife asked the mother concerning the course of the pregnancy, nutritional habits (frequency of dinners with fish or pilot whale), and use of alcohol and tobacco during the pregnancy; the answers were entered in to a questionnaire that also included information on the course of the parturition and data on the infant. Routine obstetric parameters were obtained from the medical birth registry and from the patient charts.

At age 7 years, 917 (90%) of the cohort members participated in a thorough clinical examination with focus on nervous system function (Grandjean et al., 1997). Neuropsychological tests were chosen to include tasks that would be affected by the neuropathological abnormalities described in congenital methylmercury poisoning and the functional deficits seen in children with early-life exposure to other neurotoxicants. Paper-and-pencil tests were administered by a Faroese clinical psychologist who had translated the tests into Faroese and verified their feasibility through pilot testing of Faroese children. Three computer-assisted tests were given at a separate session using the same computer.

3 Standard Analysis - confounder control

For ethical reasons randomized trials cannot be used when evaluating the (potential) adverse effects of environmental agents. Thus, in environmental epidemiology the resulting effect of the exposure is often assessed from observational data. When analyzing such data the researcher is always faced with the possibility that the exposure-response association may be confounded. A *confounder* is defined as an extraneous determinant of the response which has imbalanced distributions between the compared categories of the exposure (Miettinen, 1985). In many environmental studies, socioeconomic status is an important confounder. Subjects with a low socioeconomic status tend to be more exposed to chemical agents than others. On the other hand high socioeconomic status is typically associated with good health. Thus, if these associations are ignored then the adverse health effects of the chemical are likely to be overrated.

Thus, to obtain a correct assessment of the health effects of a given environmental substance it is important to take into account the effects of confounding variables. The first step in the confounder correction process is to identify all potential confounders. This identification should be

based on all available biological knowledge about the mechanisms being studied. In the Faroese study a set of 20 covariates including the child's sex and age, maternal Raven score (a measure of intelligence) and socioeconomic variables were identified on the basis of a priori knowledge on potential influence on the outcome variables, as considered in the light of the epidemiological setting in the Faroe Islands. Mercury exposure, which depends on local whale meat availability and personal food preferences, more than, e.g., socioeconomic factors was thought to be weakly related to most covariates. This was confirmed in multiple regression analysis showing that the confounders could not explain more than 10% of the exposure variation. However, in bivariate analysis 5 covariates were significantly (5% level) related to the exposure. Maternal intelligence and education were negatively associated with mercury exposure. Furthermore, children with a Danish mother, children in day care, and children with older siblings tended to have a lower prenatal mercury exposure. Most of these associations are a result of low consumption of whale meat in the capital of Tórshavn.

Having identified the potential confounders the aim of the preceding analysis is to estimate the effect of a given exposure increase for fixed confounder values. This can be done by stratification or through some sort of regression analysis. However, as the a priori knowledge if often limited the list of confounders may be long and the former approach is not feasible. Attention is therefore restricted to the latter approach which has the advantage of yielding a stronger analysis if the regression model is correct. In the Faroese study the full model including all potential confounders contains more than 20 nuisance parameters in addition to the parameter of interest. To gain power in the effect estimation, the standard statistical procedure prescribes identification and removal of any unnecessary covariates (Kleinbaum, Kupper and Morgenstern, 1982). In the original analysis of the Faroese data Grandjean et al. (1997) developed an *ad hoc* criterion for confounder selection combining information across different outcome variables. According to this method the child's sex and age in addition to the maternal Raven score were considered obligatory confounders for all outcome variables. For tasks performed on a computer a measure of the child's computer acquaintance was included in the set of obligatory confounders. Additional confounders were selected approximately as follows. For each neuropsychological test important predictors were identified using backward elimination (adjusted for the obligatory covariates) with p=0.10. Predictors that were important for more than 3 outcomes (out of 17) were then included in the final regression model for all outcomes.

For each of the neuropsychological outcomes Table 1 shows the estimated mercury effect using

the mercury concentration in cord blood and in maternal hair, respectively, as the exposure indicator. Exposure effects are corrected for the set of confounders identified by Grandjean et al. (1997), after inclusion of a dichotomous covariate (*Town7*) indicating whether or not the child was living in one of the three Faroese towns (Tórshavn, Klaksvik or Tvaeraa) at the time of the examination. This covariate was not considered originally but has been added to the list of confounders for control mainly because of concern that the rural children perform more poorly, perhaps also because of fatigue caused by traveling to the hospital clinic. Furthermore, rural children had a significantly ($p < 0.0001$) higher prenatal mercury exposure than urban children.

From Table 1 strong mercury effects are seen for the two Boston Naming Test (BNT) measures and the reaction time test. Furthermore, it is seen that the cord blood concentration seem to be a better predictor of childhood test performance than the concentration of mercury in maternal hair. Compared to the results of Grandjean et al. (1997) inclusion of the $Town7$ covariate tended to attenuate the mercury coefficients. In agreement with the a prior hypothesis urban children performed better than rural children on most of the tests. Thus, these results indicate that $Town7$ is an important confounder of the relationship between prenatal mercury exposure and childhood cognitive ability.

Table 1 here

4 Weaknesses of the standard analysis

In the previous section the Faroese data was analyzed using multiple regression methods. Although this approach currently is the standard method it has some serious shortcomings when applied to environmental epidemiology data.

4.1 Confounder selection uncertainty

As was the case for the Faroese study in observational epidemiology the confounders used for control are often identified based on the data. Several confounder selection methods have been suggested. Unfortunately, no standard procedure is really satisfactory. One approach (forward selection/backward elimination) is based on stepwise testing of the effects of the potential confounders on the outcome while another (change-in-estimate) removes potential confounders as long as the exposure effect does not change too much. Despite the frequent use the inferen-

tial properties of these strategies are not yet well known, and so far it has not been possible to identify an optimal procedure for confounder selection.

When the confounders have been selected the final analysis on the exposure effect is almost always conducted as if the selected model were known a priori. Thus, the uncertainty associated with the first part of the estimation process is erroneously ignored. This procedure may introduce bias in the exposure effect estimate and precision estimates are likely to be overly optimistic (Miller, 1990).

For the Faroese data Budtz-Jørgensen et al. (2002a) described how to adjust for data dependent confounder identification. Because of the complex nature of the two-stage selection estimators (first selection then estimation) no firm theory is currently available to perform such adjustments. The bootstrap method (Efron and Tibshirani, 1993) therefore constitutes the obvious choice for incorporation of model selection uncertainty in the final inference. This approach was applied to the regression analysis of Section 3. In each bootstrap sample confounders were selected and the exposure effect estimated. The statistical properties of the composite estimator were then assessed from the empirical distributions of the exposure estimates. These analyses showed that the selection method used did not underestimate the variability in the exposure coefficients by an important amount. Furthermore, the difference between effect estimates obtained in the full model and in the final model was small, indicating that the selection method is approximately unbiased. Thus, the significant mercury effects of Table 1 cannot be explained as an artifact caused by the data driven model selection process.

4.2   Exposure measurement error

The multiple regression analysis assumes that all independent variables are measured without error. This requirement is seldom satisfied for environmental exposure variables. The relevant dose is often assumed to be some sort of a long term average load. Such exposure variables cannot be measured without error. Typically, it is only feasible to measure the exposure at one or a few specific points in time. Thus, in addition to ordinary laboratory error the exposure measurement will also be subject to variations in time and biological differences between subjects.

Imprecision in the exposure variable can seriously affect the validity of the statistical analysis of the exposure effect. Thus, instead of the true (causative) exposure variable X the investigator is left with an error prone measure W. If W is naively substituted for X in the statistical analysis, then the exposure effect estimate may yield a biased reflection of the causal effect of

X. However, the statistical properties of the naive estimator depend on the model relating X to the response Y as well as the size and the nature of the measurement error (Carrol, Ruppert, Stefanski, 1995). Under the *classic additive error model* it is assumed that the exposure measurement is given as a sum of the true exposure and a random error, i.e. $W = X + U$, where the measurement error U is independent of X. The measurement error is said to be non-differential if W is independent of Y given X and the condfounders (assumed error free). In multiple regression analysis failure to adjust for non-differential measurement error will attenuate the exposure effect. The higher the imprecision the stronger this attenuation becomes. However, the attenuation also depends on the amount of confounding in the study and inference about confounder effects generally become invalid. This further complicates the task of confounder selection (Budtz-Jørgensen et al., 2002b). Furthermore, exposure measurement error increases the residual variance in the exposure-response relation which means that power to detect exposure effects is lost.

If more than one exposure indicator is available then it is possible to correct for the measurement error under certain assumptions. Budtz-Jørgensen et al. (2002b) considered two correction methods for the Faroese data. In one approach the measurement error variance of the cord blood concentration was estimated using factor analysis. This approach assumes that except for measurement error (log transformed) mercury concentrations in cord blood and maternal hair are given as linear functions of the true unobserved exposure variable. In order to obtain an identified model a third exposure indicator, number of pilot whale dinners consumed by the mother during pregnancy (log transformed), was included. Based on the estimated error variances consistent estimators of the exposure effect can be obtained using the *method of moments* (Fuller, 1987). The cord blood regression coefficients were also corrected by viewing the maternal hair concentration as a so-called *instrumental variable* (Fuller, 1987). These correction methods were in good agreement both resulting in a 15% de-attenuation of the naive mercury coefficients.

4.3 Multiple endpoints

Prospective cohort studies often include a large number of disease endpoints. A priori information about the adverse effect of the agents being studied is often weak, which makes it difficult to rule out effects in advance. Inclusion of many outcome variables will reduce the risk of overlooking important health effects. On the negative side, this will also increase the risk of chance findings, especially if each endpoint is analyzed separately without any clear a priori hypothesis.

In the Faroese study information was collected on 17 different neurobehavioural test variables. Furthermore, the prenatal mercury exposure was measured in two ways: as the concentration of mercury in maternal hair and as the mercury concentration in the cord blood. Thus, in the standard analysis of these data the hypothesis of no mercury effect was tested 34 times (Table 1). Although all associations appeared to be in the direction expected, it is not surprising that some mercury effects were found to be statistically significant at the conventional level of 5%.

Figure 1 here

Thus, to obtain a correct assessment of the significance of the estimated exposure effects it is often necessary to conduct some sort of correction for multiple testing. The Bonferroni method is the standard technique for this purpose. However, with correlated outcomes this method is known to be very conservative. This is a critical weakness when studying the relatively weak effects of low level exposure to chemical substances. In addition, the outcome variables in environmental epidemiology are often noisy which further reduces the power to detect exposure effects. The neurobehavioural outcomes in the Faroese study illustrate this point nicely. Figure 1 shows a partial residual plot of the association between prenatal mercury exposure and the scores on the cued Boston Naming Test. Even though this outcome has the strongest mercury effect this effect is seen to be quite small compared to the large residual variation.

5   Structural Equation Modeling

This section gives a general presentation of a class of statistical models which are better equipped than standard regression models to deal with the statistical challenges associated with analyzing environmental epidemiological data.

Structural equation models constitute a very general and flexible class of statistical models including ordinary regression models and factor analytic models (Bollen 1989; Arminger, Wittenberg and Schepers, 1996). The aim is to model the conditional distribution of the observed response variables $(y_i = (y_{i,1}, ..., y_{i,p})^t)$ given the observed covariates $(z_i = (z_{i,1}, ..., z_{i,q})^t)$ of subject i, $i = 1, ..., n$. This is achieved by first attaching a continuous latent response variable $y_{i,j}^*$ to each of the observed response variables. The relation between $y_{i,j}$ and $y_{i,j}^*$ depends on the nature of the observed variable. For $y_{i,j}$ continuous one simply lets $y_{i,j} = y_{i,j}^*$, while a threshold model is postulated if $y_{i,j}$ is ordered categorical or censored.

A structural equation model typically consists of two parts: a measurement model and a struc-

tural model. In the measurement model the response variable y_i is related to the covariates and a latent m-dimensional variable η_i

$$y_i^* = \nu + \Lambda \eta_i + K z_i + \epsilon_i, \tag{1}$$

where ν is a vector of intercepts, Λ is a $p \times m$ matrix of factor loadings and ϵ_i is a vector of measurement errors, which follows a normal distribution with mean zero and covariance Ω. The matrix K contains regression coefficients which describe direct effects of the covariates on the (latent) response variables. Usually only a few of the rows of K are different from zero.

The structural part of the model describes the relation between the latent variables (η_i) and the covariates

$$\eta_i = \alpha + B \eta_i + \Gamma z_i + \zeta_i \tag{2}$$

Here α is a vector of intercepts and B is an $m \times m$ matrix of regression coefficients describing the relation between the latent variables. The diagonal elements of this matrix is zero and $I - B$ is non-singular. Covariate effects are given by the $m \times q$ matrix Γ. ζ_i is an m-dimensional vector of residuals, which is assumed to be independent of the measurement errors ϵ_i, while following a normal distribution with mean zero and variance Ψ.

The model can be extended by letting some parameters depend on a group variable. For example, the parameters of the structural part of the model may depend on the gender of the subject.

5.1 Estimation

The parameters to be estimated are $\theta = (\tau, \nu, \Lambda, K, \Omega, \alpha, B, \Gamma, \Psi)$ where τ denotes the vector of all unknown thresholds. The likelihood function is derived by noting that the conditional distribution of y_i^* given z_i is $N_p\{\mu(\theta) + \Pi(\theta) z_i, \Sigma(\theta)\}$, where $\mu(\theta) = \nu + \Lambda(I - B)^{-1}\alpha$, $\Pi(\theta) = \Lambda(I - B)^{-1}\Gamma + K$ and $\Sigma(\theta) = \Lambda(I - B)^{-1}\Psi(I - B)^{-1^t}\Lambda^t + \Omega$. The model is naturally extended by letting μ, Π and Σ vary freely. The resulting model is known as *the reduced form* or *the unrestricted model* and plays a central role in the estimation algorithm for θ.

Assuming independence between subjects the likelihood function becomes $L(y, z, \theta) = \prod_{i=1}^{n} \int_{D_i} \phi(y_i^* | \mu(\theta) + \Pi(\theta) z_i, \Sigma(\theta)) dy_i^*$, where ϕ is the density of the normal distribution, and the i'th domain of integration (D_i) is the set of y_i^*-values which are mapped onto the observed value of the response y_i.

In models where all response variables are continuous the likelihood function is a product of conditional normal distribution densities, and parameters may be estimated using the maximum likelihood (ML) method. When one or more of the response variables is ordinal or censored, the likelihood function is an integral and ML estimation is currently not available in user-friendly software. Instead a weighted least squares estimation method suggested by Muthén (1984) may be used. This method consists of three steps

1. The reduced form parameters τ, μ, Π and the diagonal elements of Σ are estimated in marginal analyses of each of the p response variables $y_{i,j}\ i = 1, ..., n$. Thus, for $y_{i,j}$ continuous the estimates are obtained from an ordinary linear regression model, while an ordinal probit model is fitted if $y_{i,j}$ is ordered categorical. For identification the residual variance of categorical response variables is set to one.

2. The off diagonal elements of Σ are estimated in the bivariate distributions of all pairs of response variables. The estimates maximize the likelihood of the model for only two response variables $(y_{i,j}, y_{i,k}\ i = 1, ..., n)$ given the covariates *and* the estimates obtained in step 1.

3. Reduced form parameters are stacked in a vector κ and the parameters of the structural equation model θ are estimated by minimizing a weighted least squares discrepancy function

$$F(\theta) = \{\widehat{\kappa} - \kappa(\theta)\}^t W^{-1}\{\widehat{\kappa} - \kappa(\theta)\} \tag{3}$$

where $\widehat{\kappa}$ is the vector of estimates obtained in steps 1 and 2 and W is a weight matrix.

Different choices of weight matrix W are available in user friendly software. For the so-called WLS (weighted least squares) estimator $W = V$, where V is a consistent estimator of the asymptotic covariance matrix of $\widehat{\kappa}$ (Muthén, 1984). The (asymptotic) covariance of this estimator is estimated by evaluating

$$\widehat{\text{var}}(\widehat{\theta}_{WLS}) = n^{-1}(\Delta^t V^{-1} \Delta)^{-1} \tag{4}$$

at $\widehat{\theta}_{WLS}$, where $\Delta = \partial \kappa(\theta)/\partial \theta$.

The WLSMV (weighted least squares mean and variance adjusted) estimator uses a diagonal W matrix with estimated variances of $\widehat{\kappa}$ as elements (Muthén, Du Toit and Spisic, 1997). For this

estimator the asymptotic covariance matrix is estimated by

$$\widehat{\text{var}}(\widehat{\theta}_{WLSMV}) = n^{-1}(\Delta^t W^{-1}\Delta)^{-1}\Delta^t W^{-1}VW^{-1}\Delta(\Delta^t W^{-1}\Delta)^{-1} \tag{5}$$

Asymptotically, this estimator (WLSMV) is not as efficient as the WLS estimator. However, in simulation studies Muthén, Du Toit and Spisic (1997) found that the WLSMV estimator provides a dramatically improved performance compared to the WLS estimator. Thus, when sample sizes are moderate, inclusion of off diagonal elements in the weight matrix W seems to introduce noise rather than improving efficiency. Because of this superior performance at moderate sample sizes the WLSMV estimator is sometimes described as robust.

5.2 Test of model fit

The fit of models for normally distributed responses can be compared using ordinary likelihood ratio testing. For models where at least one of the response variables is not continuous a large sample χ^2-test of model fit (against the unrestricted model) may be obtained as $2 \cdot n \cdot F_{WLS}(\widehat{\theta}_{WLS})$, where F_{WLS} denotes the WLS discrepancy function (3). Accordingly, a large sample test comparing nested models may be obtained noting that the corresponding $2 \cdot n \cdot F_{WLS}(\widehat{\theta}_{WLS})$-difference asymptotically has a χ^2-distribution with degrees of freedom equal to the difference in dimensions between the models.

Instead of the WLS-test, Muthén, Du Toit and Spisic (1997) recommended the so-called mean and variance adjusted χ^2-test (G_{MV}), due to better statistical performance when sample sizes are moderate. This statistic is obtained as follows

$$G_{MV} = \{d^*/tr(UV)\} \cdot n \cdot F_{WLSMV}(\widehat{\theta}_{WLSMV}), \tag{6}$$

where $U = W^{-1} - W^{-1}\Delta(\Delta^t W^{-1}\Delta)^{-1})\Delta^t W^{-1}$, W is the weight matrix of the WLSMV estimator and d^* is the integer closest to $\{tr(UV)\}^2/tr\{(UV)^2\}$. This variable is approximately χ^2-distributed with d^* degrees of freedom. Unfortunately, this statistic cannot be used for comparison of two nested structural equation models since G_{MV}-differences are not χ^2-distributed.

5.3 Software

The data of the Faroese mercury study were analyzed using the statistical software packages M*plus*, version 1.01 (Muthén and Muthén, 1998) and *MECOSA 3* (Arminger, Wittenberg and

Schepers, 1996). M*plus* was preferred for the final analysis because it provides robust inferential methods, missing data analysis, user-friendly programming and high computational speed. However, this program does not allow modeling of censored response variables. The WLSMV estimator was used and the variance of these estimates was estimated using the expression (5) (default in M*plus*). Furthermore, model fit was assessed using the G_{MV}-statistic (6) unless otherwise is stated.

6   Structural equation modeling of the Faroese data

The effect of prenatal mercury exposure on childhood neuropsychological test performance is estimated in structural equation models. This approach is most powerful when the outcomes can be combined into a limited number of latent effect parameters. Therefore, the neuropsychological response variables were sorted into major nervous system functions. Budtz-Jørgensen et al. (2002c) performed a thorough analysis including most of the Faroese outcome variables. Here a preliminary analysis is presented with attention restricted to verbal outcomes. These are

Wechsler Intelligence Scale for Children - Revised Digit Spans: Digit spans of increasing length were presented until the child failed both trials in a series of the same length. The score (DS) is the total number of correct forward trials.

California Verbal Learning Test (children): A list of 12 words that can be clustered into categories was given given over five learning trails, followed by a presentation of an interference list. The child was twice requested to recall the initial list, first immediately after the presentation of the interference list and again 20 minutes later after completing some other tests. Finally, a recognition test was administered. Scores are the total number of correct responses on the learning trials ($CVLT1$), on immediate and delayed recall conditions ($CVLT2, CVLT3$) and on recognition ($CVLT4$).

Boston Naming Test: The child was presented with drawings of objects and asked to name the object. If no correct response was produced in 20 seconds a semantic cue was provided describing the type of object represented. If a correct response still was not given, a phonemic cue consisting of the first two letters in the name of the object was presented. The scores are total correct without cues ($BNT1$) and total correct after cues ($BNT2$).

The main assumption in the statistical analysis of the scores on these neuropsychological tests is that they are reflections of a underlying verbal function (η_1). Thus, each outcome is assumed to

be given as a sum of a term depending linearly on η_1 and a random error (1). To define the scale of the latent verbal function the factor loading of the response *BNT*2 is fixed at one. As a starting point measurement errors of different outcomes are assumed independent. Furthermore, these outcome variables are all modeled as continuous (conditionally) normally distributed variables as was the case in Grandjean et al. (1997) and in the standard analysis of Section 3.

The verbal function is hypothesized to be affected by the true mercury exposure (η_2). Two biomarkers of a child's prenatal mercury exposure are available: the mercury concentration in the cord blood ($B\text{-}Hg$) and the maternal hair mercury concentration ($H\text{-}Hg$). After a logarithmic transformation the relation between these variables is approximately linear. This leads to the following model for the distribution of the exposure biomarkers

$$
\begin{aligned}
\log(B\text{-}Hg) &= \nu_{B\text{-}Hg} + \lambda_{B\text{-}Hg,2} \cdot \eta_2 + \epsilon_{B\text{-}Hg} \\
\log(H\text{-}Hg) &= \nu_{H\text{-}Hg} + \lambda_{H\text{-}Hg,2} \cdot \eta_2 + \epsilon_{H\text{-}Hg}
\end{aligned}
\tag{7}
$$

where the subject index i has been suppressed for simplicity in notation. The measurement errors $\epsilon_{B\text{-}Hg}$ and $\epsilon_{H\text{-}Hg}$ are assumed to be normally distributed with means 0. Furthermore, the blood and hair measurement errors are assumed independent. Methylmercury is thought to have a biological half-life of 45 days or slightly more so the concentration present in the cord blood reflects the exposure mainly during the last couple of months of gestation. If (for instance) the true dose is some sort of a long-term average mercury concentration the assumption of independence between measurement errors in cord blood and in maternal hair may be appropriate because digested mercury is deposited in the hair with a lag time of up to 6 weeks until detectable beyond the hair root. This lag-time may ensure that the two biomarkers are not affected by the same random biological fluctuations on a temporal scale. In addition, concentrations of mercury in hair and in cord blood were determined by two different laboratories (Grandjean et al., 1992).

For identifiability it is assumed that $\lambda_{B\text{-}Hg,2} = 1$, thus the true mercury exposure has the same scale as the (log-transformed) cord blood concentration. However, even with this restriction the exposure part of model is not identified. Additional information on the prenatal mercury exposure is available from the questionnaire data on maternal nutritional habits during pregnancy. In connection with each birth a midwife asked the mother about the number of pilot whale dinners per month ($Whale$) and the number of fish dinners per week ($Fish$). The distribution of the ordered categorical variables $Whale$ (5 categories: 0,1,2,3,$\geq$4) and $Fish$ (6 categories: 0,1,2,3,4,$\geq$5) is modeled introducing latent continuous variables ($Whale^*$ and $Fish^*$) and assuming a threshold relation as described in section 5. In this example the continuous latent

variables could represent the weight of ingested whale meat and fish, respectively.

Intake of fish and pilot whale meat differ fundamentally from the measurements of mercury concentrations in hair and blood. While the latter two are determined (with a certain measurement error) by the true exposure (η_2), it may seem more natural to consider intake of fish and pilot whale meat as determinants of a true exposure: an increase in maternal whale meat intake will increase the mercury exposure, not the other way around. Bollen (1989, chapter 3) describes such response variables as *cause* indicators as opposed to the two biomarkers which enter the model as *effect* indicators. From (2) it is seen that latent variables can only be affected by the covariates and other latent variables. Thus, to incorporate $Whale^*$ and $Fish^*$ as cause indicators in the current modeling framework formally it is necessary to introduce two additional latent variables η_3 and η_4. These latent variables are identical to $Whale^*$ and $Fish^*$. Thus, the measurement model for these variables is given by $Whale^* = \eta_3$, $Fish^* = \eta_4$.

The structural part of the model is $\eta = \alpha + B\eta + \Gamma z + \zeta$ with

$$B = \begin{pmatrix} 0 & \beta_{12} & 0 & 0 \\ 0 & 0 & \beta_{23} & \beta_{24} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

The parameter β_{12} yields the effect of true mercury exposure on child verbal function. Thus, in this model the mercury effect is described using only one parameter as opposed to the ordinary regression analysis where 14 mercury coefficients were required.

Intake of whale meat and fish are assumed to affect the child's mercury exposure, but no direct effects of $Whale^*$ or $Fish^*$ on the verbal function are present in the model ($\beta_{13} = \beta_{14} = 0$). In other words true mercury exposure is considered an intermediate variable in the relation between maternal seafood intake and child verbal ability.

Potential confounders of the relation between prenatal mercury exposure and childhood test performance are included in the model as covariates. Thus, these variables are assumed to be measured without error. The confounders are allowed to be correlated with the true exposure and assumed to affect the verbal function of the child (for fixed mercury exposure).

The first component of the disturbance term $\zeta = (\zeta_1, ..., \zeta_4)^t$ models the conditional distribution of the latent verbal function given the true mercury exposure and the covariates. The second component models the distribution of the true mercury exposure given the covariates and intake

of whale meat and fish. The third and fourth component describe the conditional distribution of respectively $Whale^*$ and $Fish^*$ given the covariates. These two terms are allowed to be correlated, because the covariates are not likely to explain the strong association between intake of whale meat and fish. Figure 2 shows the socalled path diagram of the proposed structural equation model.

Figure 2 here

6.1   Results

The estimated mercury effect on the verbal test performance is $\widehat{\beta}_{12} = -1.59$ (Table 2). Thus, it is estimated, that the effect of a 10-fold increase in the (true) cord blood mercury concentration corresponds to a loss of 1.6 points on the cued BNT-test. In a two-sided test this effect is highly significant with a p-value of 0.002. The mercury effect parameter (β_{12}) is on the same scale as the cord blood mercury regression coefficient of the cued BNT-test. In Section 3 this coefficient was estimated to -1.70. Thus, the structural equation model yields approximately the same effect as the standard analysis. However, as the effect estimate of the structural equation model is corrected for measurement error it may seem a little surprising that it is a little *smaller* (numerically) than the naive regression coefficient. On the other hand, in the regression analysis the strongest exposure effect was seen for the cord blood variable on the cued BNT-test. It is not surprising that inclusion of other indicators of exposure and outcome, all showing weaker exposure effects, results in an overall effect which is weaker than the strongest individual effect.

Table 2 also shows estimated factor loadings (λ) and measurement error variances (ω^2) of the two biomarkers of prenatal mercury exposure. The quality of an indicator is not determined directly by the measurement error variances because these variances are on different scales when the factor loadings are different. The indicator with the largest error variance might be the best indicator if it also has the largest factor loading. The measurement error standard deviation of the maternal hair concentration is converted to the scale of the cord blood concentration after multiplication by the absolute value of the factor loading ratio ($\omega_{H\text{-}Hg} \cdot |\lambda_{B\text{-}Hg}/\lambda_{H\text{-}Hg}|$). From the converted error variances (Table 2) it is seen that the cord blood mercury gives the most precise reflection of true exposure. This result is in agreement with the results of Grandjean et al. (1997) and Section 3 showing that in multiple regressions the cord blood concentration was a stronger predictor of childhood cognitive deficits than the maternal hair concentration. The error variance of the cord blood indicator, corresponds to a coefficient of variation of 28%. This result

is approximately four times the documented analytical imprecision (Grandjean et al., 1992).

The results of this analysis should however be interpreted with care because the proposed model does not fit the data adequately ($\chi^2_{56} = 408$, $p = 0.0000$). Below, the model is extended by relaxing the assumption of conditional independence between verbal indicators, and by allowing the covariate effects to vary more freely.

Table 2 here

6.2 Model refinements

6.2.1 Correction for local dependence

Local dependence is present when indicators are correlated beyond what is explained by the latent constructs. In the model proposed it is assumed that a child's test scores are independent given the latent verbal level. However, it seems likely that this requirement is violated for the tests considered here. As indicated in the overview above, tests reflecting the same latent function can be collected further into subgroups in which the tasks resemble each other more. As a consequence of this additional resemblance extra correlation between related indicators is to be expected. For instance, if a child accidentally misunderstands the purpose of one of the tests then this misunderstanding is likely to be repeated when the child performs the other tests in the same subgroup. Thus, a child with a good verbal function can have relatively weak scores on all four CVLT-tests or on both the BNT-tests.

Local dependence is modeled introducing two new latent variables η_5 and η_6, which enter the model as *random effects*. In addition to the latent verbal function, the BNT-tests are assumed to depend linearly on η_5, which is normally distributed and independent of all other variables. Similarly, the CVLT-tests are assumed to depend on η_6. To be precise, the measurement model for the verbal tests is now

$$
\begin{aligned}
BNT1 &= \nu_{BNT1} &+& \lambda_{BNT1,1} \cdot \eta_1 &+& \lambda_{BNT1,5} \cdot \eta_5 &+& \epsilon_{BNT1} \\
BNT2 &= \nu_{BNT2} &+& \lambda_{BNT2,1} \cdot \eta_1 &+& \lambda_{BNT2,5} \cdot \eta_5 &+& \epsilon_{BNT2} \\
CVLT1 &= \nu_{CVLT1} &+& \lambda_{CVLT1,1} \cdot \eta_1 &+& \lambda_{CVLT1,6} \cdot \eta_6 &+& \epsilon_{CVLT1} \\
CVLT2 &= \nu_{CVLT2} &+& \lambda_{CVLT2,1} \cdot \eta_1 &+& \lambda_{CVLT2,6} \cdot \eta_6 &+& \epsilon_{CVLT2} \\
CVLT3 &= \nu_{CVLT3} &+& \lambda_{CVLT3,1} \cdot \eta_1 &+& \lambda_{CVLT3,6} \cdot \eta_6 &+& \epsilon_{CVLT3} \\
CVLT4 &= \nu_{CVLT4} &+& \lambda_{CVLT4,1} \cdot \eta_1 &+& \lambda_{CVLT4,6} \cdot \eta_6 &+& \epsilon_{CVLT4} \\
DS &= \nu_{DS} &+& \lambda_{DS,1} \cdot \eta_1 &+& &+& \epsilon_{DS}
\end{aligned}
$$

The factor loading $\lambda_{CVLT1,6}$ is fixed at 1 for identification. For the two BNT-tests, only one (additional) correlation parameter is identifiable. Here the factor loadings of both indicators on the random effect (η_5) are set to 1 while the variance of η_5 is free. Thus, the possibility of a negative correlation between the two test scores is disregarded. Local dependence could also have been introduced by freeing off-diagonal elements in Ω ($= \text{var}(y_i^*|z_i, \eta_i)$). In this way, a negative correlation between the BNT-tests could have been allowed for.

6.2.2 Correction for item bias

Under the model assumptions made so far, children on the same level of the latent verbal function are excepted of have equal test scores on each of the individual tests. If item bias (or differential response function) is present this assumption is violated. A neuropsychological test is said to be biased with respect to for instance sex, if boys tend to score consistently higher (or lower) than girls with the same ability level. In this analysis, a consequence of the assumption of no item bias is that the covariates are assumed to affect verbal indicators in the same way except for scale differences. For example, the ratio between mercury corrected regression coefficients of a given covariate on the first two CVLT-tests is equal to the ratio of the verbal function factor loadings ($\lambda_{CVLT1,1}/\lambda_{CVLT2,1}$). Comparisons of regression coefficients obtained in naive multiple regressions for each indicator suggested that the assumption of no item bias is not satisfied for the study outcomes.

Item bias is easily incorporated in the model by allowing non-zero parameters in the matrix K (1). Of course, it is not possible to identify item bias with respect to the same covariate for all indicators of a given latent variable. As a minimum one indicator has to be assumed to be unbiased. Here item bias is identified successively for the covariates. For a given covariate item bias parameters are included for all indicators expect $CVLT1$, which is assumed to unbiased. Parameters that are insignificant in successive u-tests (backward elimination) are removed from the model and a new covariate is investigated, similarly. The covariates were analyzed one at a time starting with those a priori thought to be most important (i.e. the child's age and sex and maternal intelligence). To avoid identification of spurious effects using this multiple testing procedure only parameters with a numeric u-statistic above 2.5 were considered significant.

6.3 Results of the extended model

The extended model incorporating local dependence and item bias gives a very good fit ($\chi^2_{51} = 63.3$, $p = 0.1163$). As expected none of the random effects accounting for local dependence can be ignored: in (naive) u-tests random effect variances are highly significant with test values of 5.60 (BNT) and 5.19 (CVLT). Furthermore, all random effect factor loadings are highly significant (data not shown). Inclusion of three item bias parameters improved the model fit further. Two of these parameters corresponded to item-bias caused by the child's sex.

Despite the strong improvement in model fit estimated values for the main parameters (Table 3) have changed only little as a result of incorporating local dependence and item bias. The mercury effect on the verbal function is slightly weaker but still highly significant. The measurement parameters of the mercury exposure indicators are seen to be identical to those of Table 2, indicating that these parameters are determined almost entirely by the exposure indicators and are only weakly dependent on the indicators of the verbal function.

In addition to providing a simpler presentation of the main trends in the data, the structural equation approach yielded a stronger analysis. An overall test of no exposure effects based on the multiple regression analysis of Section 3 may be obtained by assuming that the residuals of indicators are normally distributed with an unrestricted covariance matrix. The significance of the mercury effect is then assessed by testing the hypothesis that the mercury coefficient is zero for all indicators. For the cord blood indicator this test was significant with a p-value of 2.25%, while the test yielded a p-value of 27.7% for the maternal hair indicator.

Table 3 here

7 Discussion

Structural equation models were shown to be useful for interpreting complex environmental data. In this framework many of the problems commonly encountered when analyzing epidemiological data can be handled in a more satisfactory way than with standard statistical tools. Most importantly, exposure effects are easily corrected for measurement error in the exposure indicator(s) or in the confounders. This can be done either by sensitivity analysis assuming a known error size or by estimating the error variances if multiple indicators are available. By viewing the outcomes as indicators of latent variables a more parsimonious representation of the exposure effects can be obtained and power may be gained. Problems with intermediate variables, ceiling

effects in outcome distributions and multiple comparisons can also be handled using structural equation techniques. Furthermore, path diagrams appear useful to explain the main assumptions of the analysis. Thus, a carefully conducted structural equation analysis will be valuable for identification of overall trends in the data and when discussing the robustness of standard regression results to deviations from the assumptions on which they rest.

Application of the multivariate method described above may however introduce another problem. When many variables (exposures, confounders and responses) are analyzed simultaneously the subset of observations with complete data may be heavily reduced. This will decrease power but may also lead to inconsistent estimation if data are not missing completely at random. This problem was addressed in further analyses of the Faroese data (Budtz-Jørgensen et al., 2002c). Ordered categorical variables were transformed for linearity and a structural equation model consisting solely of continuous response variables was developed. For such models M*plus* allows ML estimation based on the likelihood function of all available data (complete and in-complete) under the weaker assumption that data are missing at random (Little and Rubin, 1987). This analysis yielded mercury effect estimates close to those of Table 3 (data not shown).

ACKNOWLEDGEMENTS

References

Arminger, G., Wittenberg, J. and Schepers, A. (1996). *MECOSA 3 User Guide*. Friedrichsdorf, Germay: ADDITIVE GmbH.

Bollen, K.A. (1989). *Structural Equations with Latent Variables.* John Wiley and Sons.

Budtz-Jørgensen, E., Keiding, N., Grandjean, P. (2001). Benchmark Dose Calculation from Epidemiological Data. *Biometrics* **57**, 698-706.

Budtz-Jørgensen, E., Keiding, N., Grandjean, P., Weihe P., White R.F. (2002a).Confounder Identification in Environmental Epidemiology. Assessment of Health Effects of Prenatal Mercury Exposure. To be submitted.

Budtz-Jørgensen, E., Keiding, N., Grandjean, P., Weihe P., White R.F. (2002b). Consequences of Exposure Measurement Error in Environmental Epidemiology. Submitted for publication.

Budtz-Jørgensen, E., Keiding, N., Grandjean, P., Weihe P., White R.F. (2002c).Estimation of Health Effects of Prenatal Mercury Exposure using Structural Equation Models. To be submitted.

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995). Measurement Error in Nonlinear Models, Chapman & Hall.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* London: Chapman & Hall.

Fuller, W.A. (1987). *Measurement Error Models*. Wiley.

Grandjean, P., Weihe, P., Jørgensen, P.J., Clarkson, T., Cernichiari, E. and Viderø, T. (1992). Impact of Maternal Seafood Diet on Fetal Exposure to Mercury Selenium, and Lead, *Archives of Environmental Health* **47**, 185-195.

Grandjean P., Weihe P., White R.F., Debes F., Araki S., Yokoyama K., Murata K., Sørensen N., Dahl R. and Jørgensen P.J. (1997). Cognitive Deficit in 7-Year-Old Children with Prenatal Exposure to Methylmercury, *Neurotoxicology and Teratology* **19**, 417-428.

Kleinbaum, D.G., Kupper, L.L. and Morgenstern, H. (1982). *Epidemiologic Research*. Van Nostrand Reinhold Company.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis With Missing Data.* Wiley.

Miettinen, O.S. (1985). *Theoretical Epidemiology.* Wiley.

Miller, A.J. (1990). *Subset Selection in Regression.* Chapman and Hall.

Muthén, B. (1984) A general structural model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika* **49**, 115-132.

Muthén, B., du Toit, S.H.C. and Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Accepted for publication in *Psychometrika*. Available from: www.StatModel.com.

Muthén, L.K. and Muthén, B. (1998) M*plus*. The Comprehensive Modeling Program for Applied Researchers. User's Guide. Los Angeles: Muthén & Muthén.

National Academy of Sciences (NAS) (2000). *Toxicological Effects of Methylmercury.* National Academy Press.

WHO. Methylmercury. Environmental Health Criteria 101. Geneva: World Health Organization, Geneva, 1990.

Response	Cord Blood Hg		Maternal Hair Hg	
	β	p	β	p
NES2 Finger tapping				
Preferred hand	-1.01	0.08	-1.03	0.08
Non preferred hand	-0.55	0.31	-0.91	0.11
Both hands	-1.90	0.10	-2.74	0.02
NES2 Hand-Eye Coordination				
Error score*	0.03	0.27	0.05	0.10
NES2 Continuous Performance Test				
Ln total missed*	0.22	0.07	0.08	0.52
Reaction time*	34.57	0.002	16.24	0.13
Wechsler Intelligence Scale				
Digit Spans	-0.21	0.14	-0.17	0.24
Similarities	-0.003	0.99	-0.23	0.57
Sqrt. Block Designs	-0.11	0.31	-0.06	0.59
Bender Visual Gestalt Test				
Errors on copying*	0.33	0.49	0.33	0.51
Reproduction	-0.10	0.54	0.07	0.68
Boston Naming Test				
No cues	-1.61	0.002	-1.10	0.04
With cues	-1.70	0.001	-1.12	0.03
California Verbal Learning Test				
Learning	-1.00	0.23	-0.97	0.27
Short-term repro.	-0.46	0.06	-0.41	0.11
Long-term repro.	-0.46	0.10	-0.42	0.15
Recognition	-0.26	0.21	-0.19	0.38

Table 1: Estimated effects of a 10 fold increase in mercury exposure using the cord blood mercury concentration and the mercury concentration in maternal hair, respectively, as the exposure indicator. These effects are corrected for $Town7$ in addition to the confounders identified by Grandjean et al. (1997). * Higher scores indicate an adverse effect.

Mercury Effect Parameter				*Measurement Parameters of Mercury Indicators*			
	$\widehat{\beta}$	$\widehat{s.e.}$	p	Indicator	λ	ω^2	ω^2/λ^2
Verbal Function	-1.59	0.52	0.002	$\log(B\text{-}Hg)$	1	0.015	0.015
				$\log(H\text{-}Hg)$	0.80	0.039	0.061

Table 2: Estimates of main parameters of the structural equation model. The left hand side of the table gives the estimated effect of a ten-fold increase in mercury exposure. The right hand part shows estimated factor loadings (λ), measurement error variances (ω^2) and converted variances (ω^2/λ^2) for measurements of mercury concentrations in cord blood and in maternal hair.

Mercury Effect Parameter				Measurement Parameters of Mercury Indicators			
	$\widehat{\beta}$	$\widehat{s.e.}$	p	Indicator	λ	ω^2	ω^2/λ^2
Verbal Function	-1.50	0.51	0.003	log(*B-Hg*)	1	0.015	0.015
				log(*H-Hg*)	0.80	0.039	0.061

Table 3: Estimates of main parameters in the structural equation model incorporating local dependence and item bias. The left hand side of the table gives the estimated effect of a ten-fold increase in mercury exposure. To the right estimated factor loadings (λ), measurement error variances (ω^2) and converted variances (ω^2/λ^2) for measurements of mercury concentrations in cord blood and in maternal hair are given.
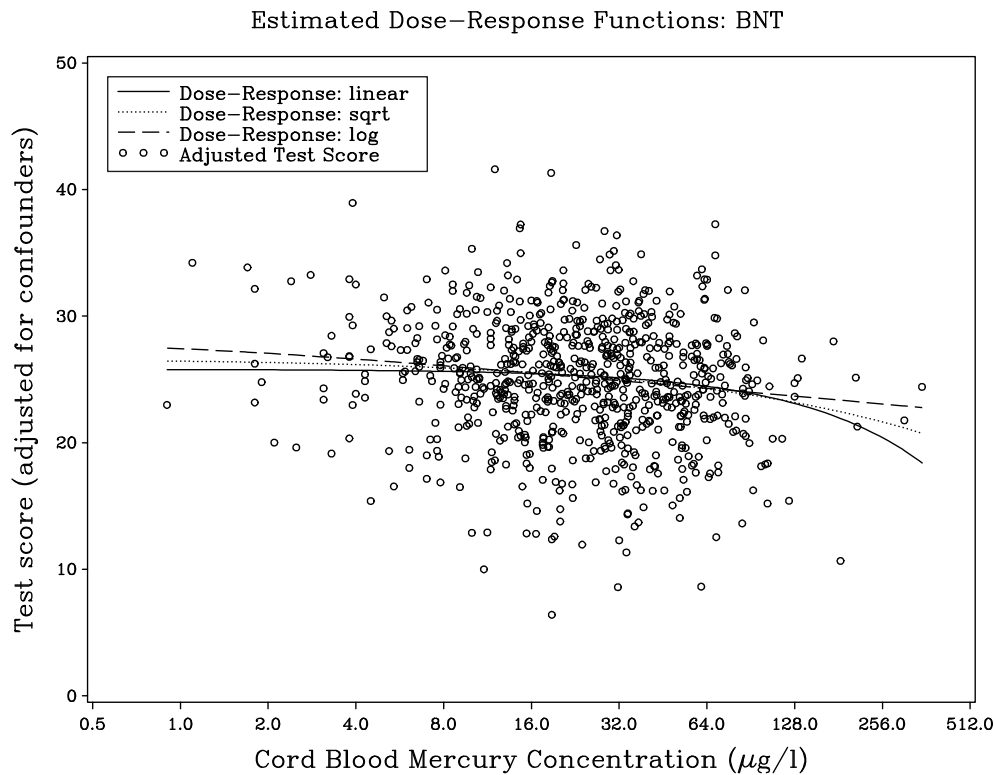

Figure 1: Partial residual plot of the relation between prenatal mercury exposure and the scores on the cued Boston Naming Test.
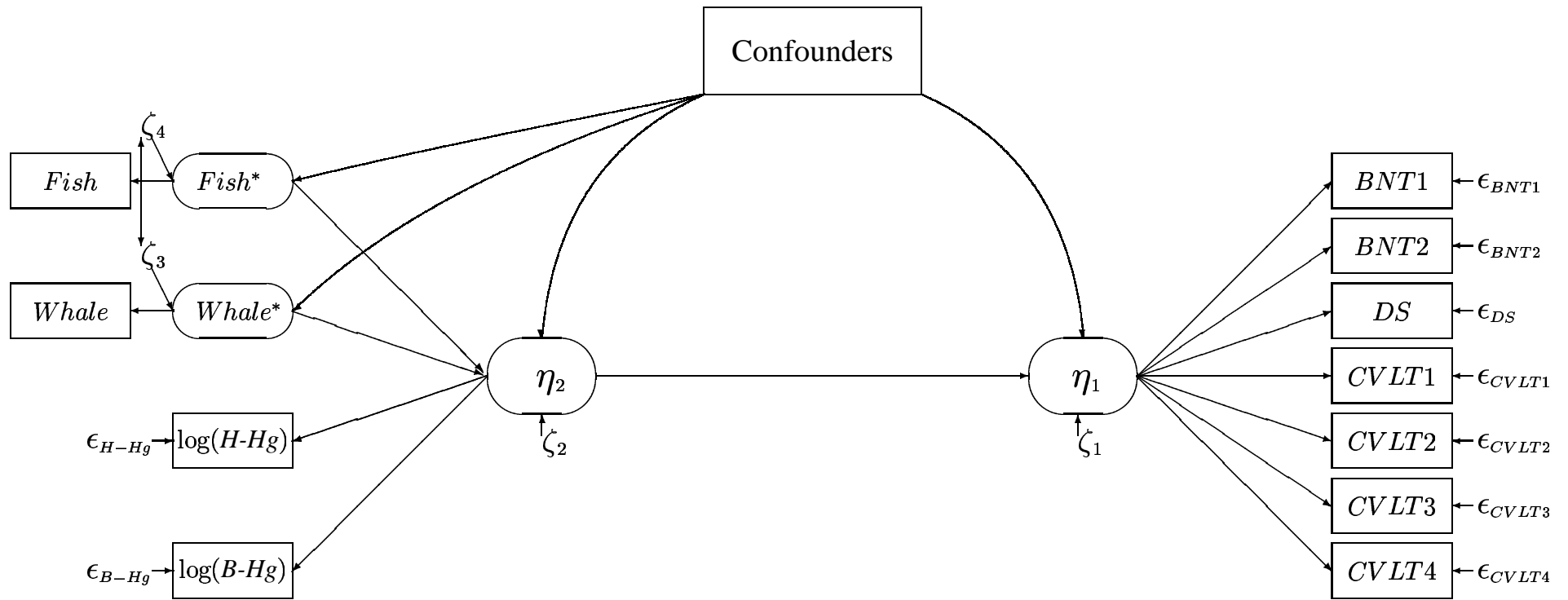
Figure 2: Path diagram for the association between indicators of mercury exposure and childhood verbal function. Here the variables η_3 and η_4 are ignored since they have been introduced only for technical reasons. In a path diagram observed variables are enclosed in boxes, latent variables are in ovals (or circles) with the exception of disturbance terms. A causal relation is represented by a single headed arrow from the causal variable to the effect variable. If two variables are connected by a two-headed arrow, this indicates that the variables are correlated but no assumptions about causation are made.