

Interpreting epidemiological evidence in the presence of multiple endpoints: an alternative analytic approach using the 9-year follow-up of the Seychelles child development study

Edwin van Wijngaarden · Gary J. Myers ·
Sally W. Thurston · Conrad F. Shamlaye ·
Philip W. Davidson

Received: 26 September 2008 / Accepted: 18 January 2009 / Published online: 11 February 2009
© Springer-Verlag 2009

Abstract

Purpose The potential for ill-informed causal inference is a major concern in published longitudinal studies evaluating impaired neurological function in children prenatally exposed to background levels of methyl mercury (MeHg). These studies evaluate a large number of developmental tests. We propose an alternative analysis strategy that reduces the number of comparisons tested in these studies. **Methods** Using data from the 9-year follow-up of 643 children in the Seychelles child development study, we grouped 18 individual endpoints into one overall ordinal

outcome variable as well as by developmental domains. Subsequently, ordinal logistic regression analyses were performed.

Results We did not find an association between prenatal MeHg exposure and developmental outcomes at 9 years of age.

Conclusion Our proposed framework is more likely to result in a balanced interpretation of a posteriori associations. In addition, this new strategy should facilitate the use of complex epidemiological data in quantitative risk assessment.

E. van Wijngaarden (✉)
Department of Community and Preventive Medicine,
University of Rochester School of Medicine and Dentistry,
601 Elmwood Avenue, Box 644, Rochester, NY 14642, USA
e-mail: edwin_van_wijngaarden@urmc.rochester.edu

G. J. Myers
Department of Neurology,
University of Rochester School of Medicine and Dentistry,
Rochester, NY, USA

G. J. Myers · P. W. Davidson
Department of Pediatrics,
University of Rochester School of Medicine and Dentistry,
Rochester, NY, USA

E. van Wijngaarden · G. J. Myers · P. W. Davidson
Department of Environmental Medicine,
University of Rochester School of Medicine and Dentistry,
Rochester, NY, USA

S. W. Thurston
Department of Biostatistics and Computational Biology,
University of Rochester School of Medicine and Dentistry,
Rochester, NY, USA

C. F. Shamlaye
Ministry of Health, Mahé, Republic of Seychelles

Keywords Methyl mercury · Child development ·
Risk assessment · Seychelles child development study

Introduction

Mercury is widespread in the environment due to multiple natural and anthropogenic sources (Counter and Buchanan 2004; McDowell et al. 2004). The organic species methyl mercury (MeHg) has been of particular concern because it occurs in low concentrations naturally and is especially toxic to the developing nervous system (WHO 1990). The primary mode of human exposure to MeHg is through fish consumption. Developmental neurotoxicity can occur following high levels of exposure to prenatal MeHg (WHO 1990). These findings raise the question of whether children whose mothers consumed fish contaminated with background levels during pregnancy are at an increased risk of impaired neurological function.

To address this important public health concern, a number of longitudinal studies have been carried out, including ones in New Zealand (Crump et al. 1998), the Seychelles islands (Myers et al. 2003) and the Faroe

Islands (Grandjean et al. 1997). From benchmark analyses of these studies, it has been estimated that a level of 4–25 parts per million (ppm) measured in maternal hair may carry a risk to the infant (Budtz-Jorgensen et al. 2000; Budtz-Jorgensen et al. 2001; Crump et al. 1998; National Research Council 2000; van Wijngaarden et al. 2006). However, risk characterization of developmental deficits in relation to prenatal MeHg exposure and its interpretation is inherently difficult due to the large number of developmental test endpoints evaluated in these studies. For example, the Seychelles child development study (SCDS) has examined subjects repeatedly and assessed 21 primary endpoints at 9 years of age (Myers et al. 2003) They have tested approximately 100 primary endpoints over the course of an 18-year follow-up. The Faroese study examined their subjects first at age 7 years and published the findings from 20 developmental tests (Grandjean et al. 1997). The New Zealand investigators based their derivation of benchmark doses on five tests scores from a larger developmental test battery of 26 tests (Crump et al. 1998).

Given this abundance of data, the question arises how to evaluate the basic hypothesis statistically. One strategy would be to analyze each endpoint individually in relation to exposure and test the statistical significance of each association. This is the approach taken in most clinical and public health research. However, under this paradigm one would expect 5% of the examined associations to be statistically significant due to random variability in the data (or “chance”). Thus, 5 out of 100 developmental tests evaluated in the SCDS in relation to prenatal MeHg exposure would be expected to be statistically significant without a guarantee that these represent true associations. Indeed, based on probability plots of p -values we previously concluded at 9 years of follow-up that the two statistically significant associations were probably chance observations, especially given the lack of statistical associations at earlier points of follow-up (Myers et al. 2003).

Some have argued that one should not rely on p -values when interpreting statistical associations, in part because systematic error such as bias and confounding may influence study findings to a much greater extent than random error (Rothman 1990; Savitz and Olshan 1995, 1998; Thompson 1998). This may be especially true in observational studies of potential health risks. Nevertheless, reliance on p -values remains common and may lead to false-positive (or false-negative) associations. Therefore, there is a need to address the analysis of multiple endpoints in relation to one primary exposure (the “multiple comparisons” issue) in order to avoid ill-informed causal inference. An often-used approach to handle such multiple comparisons is to adjust the cut-off for rejecting the null hypothesis according to the number of statistical tests performed, such as the Bonferroni adjustment (Glantz

2002). For example, in one study (or a set of longitudinal studies based on the same study population) with 100 endpoints one would divide the alpha level of 0.05 (for statistical significance at the 5% level) by 100, thus now considering only those hypothesis tests yielding a p -value of 0.0005 or less to be statistically significant. Albeit simple, this correction is overly conservative when a large number of comparisons are made (Glantz 2002), and suffers from numerous additional limitations which have been described in detail elsewhere (Perneger 1998).

A key issue in discussing the merit of multiple testing adjustments is the distinction between a priori hypotheses and a posteriori observations (Thompson 1998; Veazie 2006). Recently, more refined ways have been developed to address multiple hypotheses in a single study, including Bayesian approaches (Berry and Hochberg 1999; Gelman and Tuerlinckx 2000). A Bayesian approach to models of multiple related hypotheses may include using hierarchical models to model multiple parameters as exchangeable. This type of approach allows information about individual parameters to be influenced by the magnitude of related parameters, e.g. a borrowing of strength of information across related parameters. Bayesian or empirical Bayes methods can be appropriate for either a priori or a posteriori investigation (Efron et al. 2001; Veazie 2006). On the other hand, alternative methods for a priori investigation are not necessarily statistical in nature. Rather, they may rely on investigators specifically stating the overarching research hypothesis and determining whether this hypothesis is a composite hypothesis (i.e., a composition of single hypotheses). For instance, one could hypothesize that prenatal exposure to MeHg (or, more accurately, maternal hair levels of MeHg) is associated with a decline in “cognitive functioning” in a cohort of children, where cognitive function can be measured in a variety of different ways. Arguably, this is different from hypothesizing that prenatal MeHg exposure is specifically associated with a lower score on the Boston Naming Test, which may or may not be representative of the broader construct of cognitive functioning intended to be measured by the investigators. One might surmise that the broader construct is of true public health interest. If this is the case, it would not appear appropriate to arrive at a conclusion following from a single hypothesis reported in isolation from other hypotheses addressed in the same study. To enhance our ability to properly interpret complex data, an a priori analysis approach should be developed that reflects the composite nature of the hypothesis of interest.

The evaluation of a large number of endpoints in the MeHg literature also complicates the derivation of safe exposures. The choice of an endpoint on which to base safe exposures is often arbitrary given the variety of study populations and the methodologies employed in

epidemiological studies (van Wijngaarden et al. 2006). In the absence of conclusive evidence to suggest that subtle differences in cognitive and motor performance reported in relation to prenatal MeHg exposure impair an individual child's overall ability to function, it is uncertain which endpoint is biologically most relevant (van Wijngaarden et al. 2006). Historically, some agencies have used one single endpoint from one single study to derive a safe MeHg exposure (National Research Council 2000), whereas other researchers have aggregated the original data from three prospective studies to evaluate the impact on children's intelligence quotient (IQ) (Axelrad et al. 2007; Cohen et al. 2005). We recently suggested presenting an average benchmark dose and its corresponding range based on all available evidence to provide an indication of the exposure limits within which the true benchmark dose is likely to fall (van Wijngaarden et al. 2006).

One alternative to evaluating a large number of endpoints separately is to fit a single model that includes information on all outcomes simultaneously. Two different approaches have been taken in the statistical literature to do this, and both have been applied to MeHg exposure. Both approaches allow the outcomes to cluster into different functions or "domains", such as "cognitive functioning", "motor functioning", etc., and both allow estimation of the function-specific effects. One approach uses structural equations models, which introduce a latent variable for each function or domain (Budtz-Jørgensen et al. 2002). The second approach models the MeHg effect directly, by including an overall MeHg effect, and domain-specific and outcome-specific deviations (Thurston et al. 2009).

In this study, we propose another alternative method to reduce the number of statistical comparisons made in the evaluation of the risk of developmental deficits in relation to prenatal MeHg exposure. Whereas most previous studies have examined results of individual developmental tests in isolation, our definition of 'abnormal' versus 'normal' performance integrates information from all administered developmental tests into one ordinal endpoint. This approach drastically reduces the number of endpoints to be evaluated in statistical analyses. It is based on the premise that no one endpoint is biologically more relevant than any other, and that a deficit in any of the tests may be important in later life. Additionally, the nature of brain damage due to high doses of prenatal MeHg suggests that cellular functions associated with very basic brain development are impacted (Clarkson 2002). Therefore, one would expect a similarity of prenatal MeHg exposure effects on endpoints in different domains (Thurston et al. 2009). Indeed, the multiple outcomes models used by Thurston and colleagues as applied to these data suggested a similarity of exposure effects on all outcomes, when the outcomes were put on a comparable scale. Most previous reports relied on

generalized linear models to evaluate the association between prenatal MeHg exposure and child development, the results of which may be difficult to interpret given the subtle effects reported. Other measures of effect, such as the relative risk, may allow for a more intuitive interpretation of findings. We computed the relative risk for an adverse event and the corresponding 95% confidence interval across the range of prenatal MeHg exposure observed in the SCDS.

Materials and methods

Study population

The SCDS Main cohort has been extensively described previously and is briefly summarized here (Davidson et al. 1998; Myers et al. 2003; Shamlaye et al. 2004). In 1989–1990, we initiated a longitudinal study of a cohort of 779 mother–child pairs in the Seychelles Islands (approximately half of the number of live births during that period). The children were 6 months \pm 2 weeks old at enrollment. The primary purpose of this study was to examine the association between prenatal MeHg exposure from maternal fish consumption during pregnancy and the children's developmental outcomes. Most Seychellois consume ocean fish with 10 or more meals each week and the fish in Seychelles contains background levels of MeHg similar to fish sold commercially in the United States. Prenatal exposure to MeHg was determined by measuring total mercury in maternal hair during pregnancy in the sample that best recapitulated prenatal exposure. We used cold vapor atomic absorption spectroscopy with quality control procedures and assumed a hair growth rate of 1.1 cm per month with a delay of 20 days between current blood concentrations and appearance of Hg in the first centimeter of scalp hair (Cernichiari et al. 1995). Postnatal exposure was measured in the 1 cm of hair closest to the scalp taken at the time of the evaluation. The study is double blind and individual MeHg exposure levels are not shared with clinical investigators or anyone in Seychelles. A variety of covariates known to be related to child development and believed to be potential confounders of the association between prenatal MeHg exposure and development were measured. These included maternal intelligence, evaluation of the home environment using the HOME, and the family's socioeconomic status. A more detailed description of the MeHg exposure assessment and collection of covariate information is available elsewhere (Crump et al. 2000; Myers et al. 2003).

A small number of children enrolled in the study have been excluded over time for illnesses or injuries known to be associated with developmental deficits, including children with closed head trauma and meningitis. Seven

hundred and seventeen children (92% of the initial cohort; 62 exclusions) were still eligible to participate at 9 years of age (Myers et al. 2003). Another 74 children were not tested at 9 years because they were residing abroad, refused to participate, or we were unable to locate them. Enrolled children were similar to those not participating (Myers et al. 2003). Thus, the current analysis is based on a cohort of 643 children (83% of the original cohort) followed for 9 years. For some analyses, we excluded up to 144 additional children due to missing outcome and covariate data (see Statistical Analysis section, and footnote to Table 3).

Developmental assessment

The children underwent extensive evaluations when they were 6, 19, 29, 66, and 107 months of age (Shamlaye et al. 2004). Our current analysis focuses on the examinations performed at 9 years (107 months) of age (Myers et al. 2003). Developmental evaluations at that age included tests for three general domains (tests for cognition and achievement; tests for motor, perceptual motor and memory; and tests for behavior) (Table 1). A team of three specially trained Seychellois child health and development professionals conducted the examinations. Reliability between the testers and the senior psychologist on the team was conducted regularly and was high.

For this analysis we initially considered the original 21 primary endpoints reported on previously (Myers et al. 2003) and seven additional secondary endpoints determined at that age: the Wechsler Intelligence Scale for Children III (WISC-III) verbal and performance IQ; WISC-III similarities and digit span subtests; the California Verbal Learning Test's (CVLT) immediate recall and recognition hits; and the Boston Naming Test's (BNT) total correct with and without cues (van Wijngaarden et al. 2006). To be consistent with our recent benchmark analysis (van Wijngaarden et al. 2006), two continuous performance task (CPT) endpoints previously reported (Myers et al. 2003) were not included; risk-taking (direction of adverse effect is unclear) and hit reaction time (statistical estimation algorithm did not converge in previous analyses). We then excluded five endpoints that had missing data for 5% or more (i.e., $n \geq 32$) of all children in the database, and also excluded three IQ measures which overlap with the constructs that are measured by the verbal and performance IQ measures.

Case definition

For our current analyses, we defined cases as those with an abnormal score on at least one of the 18 remaining developmental endpoints considered. We defined abnormal

scores using the following algorithm. We determined the 1st or 99th percentile (depending on the direction of the adverse effect) of the distribution in the cohort to identify children with an abnormal score for each subtest based on these cut-offs (Table 1). To evaluate the sensitivity of our findings to the cut-point chosen to identify children with an abnormal score, we also determined the 5th and 95th percentile (Table 1). Finally, children were classified into three case groups based on the number of tests on which they had an abnormal score: no abnormal test score on any of the 18 endpoints, one endpoint with an abnormal score, and two or more endpoints with an abnormal score. Thus, our case definition was ordinal in nature and was preserved in our statistical analyses. Case status was defined in this manner considering all endpoints combined, as well as grouped in three developmental domains consistent with our previous analyses of these data estimating benchmark dose levels (van Wijngaarden et al. 2006): cognition; motor, perceptual motor, and memory; and behavior (Table 1).

Statistical analysis

The odds ratio (OR) and 95% confidence interval (CI) for the association between prenatal MeHg level and abnormal development (as defined above, using the case definition defined by the 1st and 99th percentiles as the primary outcome) was estimated using ordinal logistic regression (proportional odds models), which is more efficient than binary logistic regression and produces risk estimates which can be interpreted across multiple dichotomizations of the outcome (Ananth and Kleinbaum 1997; Scott et al. 1997). The OR estimated with this model can be interpreted as the odds of more severe responses to less severe responses (e.g., 2+ abnormal test scores vs. 1 and 0 (combined) abnormal tests scores, or 1 and 2+ abnormal test scores (combined) vs. 0 abnormal values) among those with greater MeHg exposure as compared to the odds among those with lower levels of MeHg exposure. That is, the proportional odds model simultaneously compares children with 2+ abnormal outcomes to children with 0 or 1, and compares children with 1+ abnormal outcomes to those with no abnormal outcomes. The intercept for each of these three categories is allowed to differ, but the model assumes a common slope for MeHg for each of the two comparisons. All analyses were performed based on case status for all endpoints combined, as well as for cognition and motor function separately. We did not perform separate analyses for behavior because this domain included only one developmental test (Child Behavior Checklist) and, by definition, the number of children with an abnormal score on this test was small ($n = 7$, 1.1%) for the primary outcome based on the 99th percentile. We repeated the analyses on all endpoints combined and cognitive and

Table 1 Cut-off scores on developmental tests used to identify cases in the Seychelles child development study main cohort at 9 years of follow-up

Test	Number missing	1st or 99th percentile		5th or 95th percentile	
		Cut-off score	Number of cases (%)	Cut-off score	Number of cases (%)
Cognition					
WISC-III					
Verbal IQ	0	≤55	9 (1.4)	≤63	35 (5.4)
Performance IQ	0	≤53	7 (1.1)	≤63	35 (5.4)
California verbal learning test					
Short delay recall (scaled −3.5 to 2.5 in 0.5 steps)	1	≤−3.0	9 (1.4)	≤−2.0	38 (5.9)
Immediate Recall (scaled −4.5 to 2.5 in 0.5 steps)	1	≤−3.0	7 (1.1)	≤−1.5	54 (8.4)
Recognition hits (scaled −5 to 1 in 0.5 steps)	3	≤−3.0	11 (1.7)	≤−1.5	46 (7.2)
Boston naming test					
Total correct, no cues (scored 12–38)	6	≤13	7 (1.1)	≤16	54 (8.4)
Total correct, with cues (scored 0–12)	6	<1	6 (0.9)	<1	6 (0.9)
Motor, perceptual motor, and memory					
Bruininks–Oseretsky test of motor development	11	≤34	7 (1.1)	≤42	32 (5.0)
Berry-Buktenica VMI	9	≤69	8 (1.2)	≤78	32 (5.0)
Wide range assessment of memory and learning					
Design memory subtest (scaled 1–17)	0	≤1	8 (1.2)	≤3	64 (10.0)
Trail making					
Time to complete—trial A	4	≥100	7 (1.1)	≥63	34 (5.3)
Time to complete—trial B	20	≥258	7 (1.1)	≥189	32 (5.0)
Finger tapping					
Number of taps, preferred hand	1	≤22.4	7 (1.1)	≤25.2	33 (5.1)
Number of taps, non-preferred hand	1	≤21.6	9 (1.4)	≤23.2	35 (5.4)
Grooved pegboard					
Time to complete (in seconds), preferred hand	0	≥159	7 (1.1)	≥126	33 (5.1)
Time to complete (in seconds), non-preferred hand	2	≥204	8 (1.2)	≥146	34 (5.3)
Haptic free-form solids discrimination test—total correct (scored 0–10)	1	≤0	10 (1.6)	≤1	46 (7.2)
Behavior					
Child behavior checklist	6	≥58	7 (1.1)	≥54	40 (6.2)
Not included in case definition					
WISC-III					
Full scale IQ	0	Overlap with verbal and performance			
Similarities subtest (scaled 1–14)	0	Overlap with verbal			
Digit span subtest (scaled 1–17)	0	Overlap with verbal			
California verbal learning test					
Long delay recall (scaled −3.5 to 2.5 in 0.5 steps)	40	Missing data			
Woodcock–Johnson achievement test					
Letter-word subtest	40	Missing data			
Applied problems subtest	35	Missing data			
Connors teacher rating scale—cognitive problems	98	Missing data			
Connors continuous performance task—attentiveness (<i>d'</i>)	110	Missing data			

motor function separately based on the 5th and 95th percentile to define case status in order to evaluate the sensitivity of our findings to the cut-point chosen.

The relationship with prenatal MeHg exposure was examined using prenatal MeHg exposure as a continuous variable (as has been our practice in all previous analyses

of the SCDS data) and also in a separate model using categories of MeHg exposure based on quartiles of the exposure distribution in the cohort of 643 children. All ORs for both models were adjusted for the effects of gender, family status (categorical), child's age at testing (continuous), HOME (categorical), caregiver's IQ (continuous), Hollingshead SES (categorical), and postnatal MeHg exposure (continuous). Other covariates were not included because they did not contribute meaningfully to the prediction of the outcome or missing data reduced the sample size significantly.

For the complete case analysis, we excluded children with missing data on any of the covariates included in the model ($n = 107$) from the analysis. We compared unadjusted and adjusted OR estimates to assess the extent of confounding. However, excluding children with missing covariate data may impact the findings if the children who were excluded have attributes different from those upon which the model was based, or if the association of interest is different for the two different groups of children. To examine the potential bias introduced by the complete case approach, we compared the results from an unadjusted ordinal logistic regression before and after excluding children with missing covariate data. We assumed that ordinal logistic regression was an appropriate choice for modeling our data since the Score Test for the Proportional Odds assumption was not statistically significant ($p > 0.05$) for any of our regression models (Stokes et al. 2000). All analyses were conducted with SAS (SAS Institute, Cary, NC).

The analysis of prenatal MeHg exposure as a continuous variable described above assumes a multiplicative relationship between exposure and the risk for developing the adverse outcome of interest (e.g., OR = 1.1 for one unit increase in exposure, and OR = $1.1 \times 1.1 = 1.21$ for two units increase in exposure). However, alternative forms of characterizing the exposure-response relationship may also be informative and may allow for a more straightforward estimation of safe levels of exposure (see "Discussion"). Therefore, we additionally derived a linear slope estimate by fitting a trend line for the underlying linear relative risk (RR) model ($RR = 1 + \beta \times x$) to the ORs obtained from the models fit previously which treated exposure as a categorical variable (Rothman 1986; van Wijngaarden and Hertz-Picciotto 2004). The trend line is estimated using the median exposure concentrations and relative risk estimate in each of the exposure categories, and weights computed as the inverse variance of the relative risks (Rothman 1986). In this linear model, β represents the linear slope estimate using categorized prenatal MeHg exposure and x is the exposure level in units of ppm in maternal hair. Thus, β can be interpreted as the increase in relative risk of more severe responses to less severe responses per ppm of prenatal MeHg exposure. The lowest exposure group is used

as the reference category and is not included in the computations since no variance can be estimated. In summary, using this linear model a straight line (trend line) is fit through the category-specific relative risks while forced through the origin (baseline exposure and RR of 1.0). Estimates of the linear slope were obtained with Microsoft Excel (Microsoft Corp, Redmond, WA).

Results

Fifty-three (8.2%) children could not be classified according to overall case status due to missing data on one or more endpoints. Nine (1.4%) and 40 (6.2%) children could not be classified according to case status for cognition and motor function, respectively. Of the children with complete information on all 18 endpoints, 65 children (11.0%) scored abnormally on one or more endpoints based on cut-points using the 1st or 99th percentile, 20 (3.4%) of whom scored abnormally on two or more endpoints. Regarding cognition endpoints, 39 (6.2%) of the children with complete data had one or more abnormal test scores, and 11 (1.7%) scored abnormally on two or more endpoints. On motor function tests, 47 (7.8%) and 10 (1.7%) scored abnormally once or at least twice, respectively. Finally, 7 children (1.1%) scored abnormally on the one test representing behavior. The number of cases increased substantially when the 5th or 95th percentile was chosen to identify abnormal scores (Table 1).

Table 2 presents the association between developmental status and selected covariates. Boys were slightly more likely to do worse than girls on all tests, as were children with no biological parents in the home. Children with a higher score on HOME, higher socioeconomic status, and higher caregiver's IQ were more likely to perform better on the developmental tests. Older children performed better overall and on motor function tests, but not on tests of cognition. Postnatal MeHg exposure measured at 9 years of age did not appear to impact performance. Findings were similar for case status based on 1st/99th and 5th/95th percentiles with the exception of family status which did not remain associated with cognition case status when based on 5th or 95th percentiles.

The results for the primary outcome from the ordinal logistic regression analyses are shown in Table 3. Prenatal MeHg exposure was not associated with developmental outcomes before or after adjustment for covariates in complete case analyses. Based on the 1st and 99th percentiles, results are compatible with both an approximately 8% decrease and increase (based on the lower and upper 95% confidence limits of the risk estimates, respectively) in worse developmental performance per 1 ppm MeHg in maternal hair. After adding participants who were excluded

Table 2 Relationship between covariates and developmental case status: ordinal logistic regression without control for covariates^a

Covariate	Total cases		Cognition cases		Motor function cases	
	$N_0; N_1; N_2$	OR (95% CI)	$N_0; N_1; N_2$	OR (95% CI)	$N_0; N_1; N_2$	OR (95% CI)
Developmental case status based on 1st and 99th percentiles						
Sex (male or female)						
Girl	273; 22; 11	1.00 (reference)	306; 14; 4	1.00 (reference)	289; 16; 6	1.00 (reference)
Boy	243; 23; 9	1.37 (0.84–2.24)	289; 14; 7	1.24 (0.65–2.38)	267; 21; 4	1.22 (0.67–2.21)
Family status (no. of biological parents in home)						
2	288; 27; 8	1.00 (reference)	328; 11; 3	1.00 (reference)	305; 19; 6	1.00 (reference)
1	204; 24; 10	1.38 (0.84–2.29)	239; 15; 6	2.07 (1.03–4.15)	225; 16; 3	1.02 (0.55–1.90)
None	24; 3; 2	1.77 (0.64–1.87)	28; 2; 1	2.53 (0.69–9.30)	26; 2; 1	1.42 (0.41–5.00)
Child's age at testing (continuous)	–	0.55 (0.25–1.19)	–	2.98 (1.08–8.19)	–	0.20 (0.07–0.54)
Score for home observation for measurement of the environment (HOME)						
≤31	164; 23; 13	1.00 (reference)	198; 15; 8	1.00 (reference)	181; 16; 6	1.00 (reference)
31–35	161; 15; 2	0.47 (0.25–0.86)	182; 6; 1	0.33 (0.14–0.78)	171; 11; 1	0.57 (0.27–1.18)
>35	169; 12; 3	0.40 (0.21–0.75)	186; 7; 1	0.37 (0.16–0.84)	180; 7; 1	0.36 (0.16–0.83)
Caregiver intelligence quotient (continuous)	–	0.99 (0.97–1.01)	–	0.97 (0.95–1.00)	–	0.99 (0.96–1.01)
Hollingshead socioeconomic status						
Unskilled	181; 25; 11	1.00 (reference)	215; 12; 9	1.00 (reference)	197; 18; 6	1.00 (reference)
Semiskilled	161; 13; 5	0.56 (0.31–1.02)	187; 8; 1	0.48 (0.22–1.08)	175; 8; 2	0.47 (0.22–1.00)
Skilled; or minor/major business/profession	164; 15; 4	0.58 (0.32–1.04)	182; 8; 0	0.44 (0.19–1.01)	174; 10; 2	0.56 (0.27–1.16)
Postnatal MeHg at 107 months (continuous)	–	1.02 (0.94–1.10)	–	0.97 (0.87–1.09)	–	1.06 (0.97–1.16)
Developmental Case Status Based on 5th and 95th percentiles						
Sex (male or female)						
Girl	167; 72; 67	1.00 (reference)	246; 50; 28	1.00 (reference)	211; 59; 41	1.00 (reference)
Boy	124; 93; 67	1.39 (1.03–1.89)	224; 47; 39	1.25 (0.88–1.77)	185; 80; 27	1.12 (0.80–1.56)
Family status (no. of biological parents in home)						
2	169; 85; 69	1.00 (reference)	261; 51; 30	1.00 (reference)	224; 71; 35	1.00 (reference)
1	108; 71; 59	1.29 (0.94–1.76)	186; 42; 32	1.30 (0.91–1.87)	152; 63; 29	1.26 (0.89–1.77)
None	14; 9; 6	1.11 (0.54–2.26)	23; 4; 4	1.17 (0.51–2.67)	20; 5; 4	1.00 (0.45–2.23)
Child's age at testing (continuous)	–	0.89 (0.55–1.44)	–	2.06 (1.19–3.59)	–	0.44 (0.26–0.74)
Score for home observation for measurement of the environment (HOME)						
≤31	81; 65; 54	1.00 (reference)	144; 38; 38	1.00 (reference)	124; 50; 29	1.00 (reference)
31–35	91; 49; 38	0.68 (0.47–0.99)	141; 33; 15	0.59 (0.39–0.91)	121; 46; 16	0.77 (0.51–1.15)
>35	108; 44; 32	0.51 (0.35–0.74)	160; 25; 9	0.37 (0.24–0.59)	138; 35; 15	0.56 (0.37–0.85)
Caregiver intelligence quotient (continuous)	–	0.99 (0.98–1.00)	–	0.98 (0.96–0.99)	–	0.99 (0.98–1.00)
Hollingshead socioeconomic status						
Unskilled	92; 66; 59	1.00 (reference)	157; 42; 37	1.00 (reference)	137; 52; 32	1.00 (reference)
Semiskilled	89; 46; 44	0.79 (0.54–1.14)	151; 25; 20	0.59 (0.39–0.90)	121; 44; 20	0.84 (0.56–1.24)
Skilled; or minor/major business/profession	105; 49; 29	0.54 (0.37–0.79)	154; 28; 8	0.44 (0.28–0.69)	132; 40; 14	0.64 (0.42–0.96)
Postnatal MeHg at 107 months (continuous)	–	1.02 (0.97–1.07)	–	1.01 (0.95–1.07)	–	1.01 (0.96–1.07)

N_0 number of children with no abnormal test score, N_1 number of children with 1 endpoint with an abnormal test score, N_2 number of children with two or more endpoints with an abnormal test score

^a None of the models violated the score test for the proportional odds assumption; “behavior” cases not examined separately

from complete case analyses due to missing data on covariates, risk estimates for the continuous exposure measure were quite similar although point estimates for the categorical analyses were somewhat lower. Nevertheless, we considered them qualitatively comparable due to

overlapping confidence intervals. The interpretation of the findings did not change substantially across case definitions (based on 1st/99th or 5th/95th percentiles), with the possible exception of slight inverse association in the linear relative risk model (Table 3).

Table 3 Developmental case status in relation to prenatal MeHg level: ordinal logistic regression^a

Prenatal MeHg level (ppm) ^b	Total $N_0; N_1; N_2$	Unadjusted logistic regression ^c		Multiple logistic regression ^c		Unadjusted logistic regression ^d	
		OR	95% CI	OR	95% CI	OR	95% CI
Developmental case status based on 1st and 99th percentiles							
Total cases							
≤3.34	124; 14; 6	1.00	(reference)	1.00	(reference)	1.00	(reference)
>3.34–≤5.94	137; 10; 6	1.03	(0.46–2.28)	1.10	(0.48–2.52)	0.73	(0.36–1.47)
>5.94–≤9.28	127; 14; 4	1.22	(0.56–2.66)	1.18	(0.52–2.65)	0.87	(0.44–1.72)
>9.28	128; 16; 4	1.03	(0.46–2.28)	1.00	(0.44–2.29)	0.95	(0.49–1.86)
Continuous	–	1.004	(0.943–1.069)	1.004	(0.940–1.073)	1.009	(0.957–1.064)
Continuous (linear β)	–	0.0081	(–0.048 to 0.064)	0.0067	(–0.050 to 0.063)	–0.015	(–0.055 to 0.025)
Cognition cases							
≤3.34	146; 7; 5	1.00	(reference)	1.00	(reference)	1.00	(reference)
>3.34–≤5.94	151; 5; 3	0.63	(0.20–1.97)	0.70	(0.21–2.35)	0.65	(0.26–1.64)
>5.94–≤9.28	149; 8; 1	1.02	(0.37–2.79)	1.20	(0.40–3.58)	0.75	(0.31–1.83)
>9.28	149; 8; 2	1.00	(0.36–2.75)	1.05	(0.36–3.08)	0.83	(0.35–1.98)
Continuous	–	0.997	(0.916–1.086)	1.003	(0.916–1.098)	0.989	(0.919–1.064)
Continuous (linear β)	–	–0.013	(–0.077 to 0.051)	–0.0033	(–0.077 to 0.071)	–0.027	(–0.072 to 0.019)
Motor function cases							
≤3.34	135; 10; 3	1.00	(reference)	1.00	(reference)	1.00	(reference)
>3.34–≤5.94	144; 8; 4	1.57	(0.59–4.19)	1.57	(0.57–4.29)	0.87	(0.39–1.98)
>5.94–≤9.28	137; 11; 1	1.56	(0.59–4.16)	1.38	(0.50–3.80)	0.90	(0.40–2.04)
>9.28	140; 8; 2	0.84	(0.27–2.56)	0.78	(0.25–2.43)	0.74	(0.31–1.74)
Continuous	–	0.995	(0.919–1.077)	0.992	(0.912–1.079)	0.995	(0.931–1.064)
Continuous (linear β)	–	0.0034	(–0.066 to 0.073)	–0.0039	(–0.070 to 0.062)	–0.020	(–0.064 to 0.023)
Developmental case status based on 5th and 95th percentiles							
Total cases							
≤3.34	69; 40; 35	1.00	(reference)	1.00	(reference)	1.00	(reference)
> 3.34–≤5.94	73; 50; 30	0.79	(0.49–1.27)	0.78	(0.48–1.26)	0.77	(0.50–1.19)
>5.94–≤9.28	74; 41; 30	0.86	(0.53–1.38)	0.89	(0.55–1.44)	0.90	(0.58–1.39)
>9.28	75; 34; 39	1.33	(0.83–2.14)	1.32	(0.82–2.15)	1.20	(0.78–1.85)
Continuous	–	1.034	(0.995–1.074)	1.034	(0.994–1.074)	1.025	(0.991–1.060)
Continuous (linear β)	–	–0.0037	(–0.037 to 0.030)	–0.0029	(–0.037 to 0.032)	–0.0048	(–0.035 to 0.025)
Cognition cases							
≤3.34	118; 19; 21	1.00	(reference)	1.00	(reference)	1.00	(reference)
>3.34–≤5.94	119; 31; 9	0.46	(0.26–0.81)	0.45	(0.25–0.79)	0.51	(0.31–0.84)
>5.94–≤9.28	114; 28; 16	0.62	(0.35–1.09)	0.65	(0.36–1.14)	0.66	(0.40–1.10)
>9.28	119; 19; 21	1.08	(0.61–1.90)	1.12	(0.63–1.99)	1.00	(0.60–1.66)
Continuous	–	1.013	(0.969–1.059)	1.016	(0.971–1.063)	1.009	(0.971–1.049)
Continuous (linear β)	–	–0.049	(–0.077 to –0.020)	–0.050	(–0.079 to 0.020)	–0.041	(–0.067 to –0.014)
Motor function cases							
≤3.34	91; 35; 22	1.00	(reference)	1.00	(reference)	1.00	(reference)
>3.34–≤5.94	108; 29; 19	0.88	(0.52–1.47)	0.94	(0.55–1.61)	0.72	(0.45–1.14)
>5.94–≤9.28	95; 40; 14	1.00	(0.60–1.67)	1.03	(0.61–1.75)	0.85	(0.54–1.35)
>9.28	102; 35; 13	0.88	(0.52–1.48)	0.90	(0.53–1.54)	0.72	(0.45–1.14)
Continuous	–	1.001	(0.960–1.044)	1.005	(0.962–1.049)	0.986	(0.951–1.024)

Table 3 continued

Prenatal MeHg level (ppm) ^b	Total N_0 ; N_1 ; N_2	Unadjusted logistic regression ^c		Multiple logistic regression ^c		Unadjusted logistic regression ^d	
		OR	95% CI	OR	95% CI	OR	95% CI
Continuous (linear β)	–	–0.0093	(–0.040 to 0.022)	–0.0060	(–0.039 to 0.027)	–0.026	(–0.049 to –0.0034)

N_0 number of children with no abnormal test score, N_1 number of children with 1 endpoint with an abnormal test score, N_2 number of children with two or more endpoints with an abnormal test score

^a None of the models violated the score test for the proportional odds assumption

^b Median exposure level; ≤ 3.34 : 2.2 ppm; > 3.34 – ≤ 5.94 : 4.7 ppm; > 5.94 – ≤ 9.28 : 7.6 ppm; > 9.28 : 12.3 ppm

^c Adjusted for the effects of gender, family status code, score for HOME observation, Hollingshead socioeconomic status (all categorical); average age of child at testing, caregiver intelligence quotient, and postnatal MeHg level (all continuous); all results are based on *complete case analysis*, i.e., excluding children with missing data on any of the covariates in the model ($n = 107$) in addition to those with missing outcome data; accordingly, the total number of children used in the analyses were 499 for the total case analysis, 528 for the cognition analysis, and 508 for the motor function analysis

^d Unadjusted analyses, including all children regardless of missing covariate data—only children with missing outcome data excluded from analyses ($n = 53$ for total cases, $n = 9$ for cognition cases, and $n = 40$ for motor function cases)

Discussion

In our present analysis of the SCDS at 9 years of follow-up, we chose to combine the test results a priori for 18 individual endpoints into one overall ordinal outcome variable, and two ordinal outcome variables representing functional domains. In this analysis there was no significant association between prenatal MeHg exposure and developmental functioning at 9 years of age. This is consistent with our previous reports using separate models for each endpoint (Myers et al. 2003). However, we argue that this new strategy is more likely to result in a balanced interpretation of a posteriori associations, because those reviewing and utilizing the findings are not distracted by statistically significant results for individual endpoints of unknown statistical and clinical value.

Our findings were expressed in odds ratios, which are perhaps easier to interpret by policy makers, clinicians and public health investigators than estimates obtained from linear regression or more broadly generalized linear models. Despite the lack of any consistent associations, our data still provide information that can be used to guide regulatory decision-making. Odds ratios reported for a one unit increase in exposure can be used for risk assessment, i.e., to estimate safe levels of exposure using methods similar to those reported previously for quantitative cancer risk assessment (van Wijngaarden and Hertz-Picciotto 2004). The maternal hair MeHg level in ppm required to produce a 1% additional risk of morbidity over the background risk (here called “hazardous dose”, or HD) can be calculated by first computing the relative risk corresponding to this additional risk using the following equation:

$$RR_{HD1} = \frac{R(0) + 0.01}{R(0)}$$

where $R(0)$ is the background risk of morbidity. Subsequently, this RR_{HD1} is substituted into the following equation:

$$HD1 = \frac{RR_{HD1} - 1}{\beta} \text{ (derived from } RR_{HD1} = 1 + \beta \times HD1 \text{)}$$

where β is the estimated slope from the linear relative risk model. Given our ordinal outcome data, we can compute two types of HD1 estimates. For example, the background risk $R(0)$ of an abnormal score on 1 or more tests in the reference group (i.e., ≤ 3.34 ppm based on quartiles of the exposure distribution in this cohort) for “total cases” is $20/144 = 13.9\%$ (see Table 3). That is, of the 144 children exposed to less than 3.34 ppm prenatal MeHg, 14 had one endpoint (N_1) and six had two or more endpoints with an abnormal score (N_2). The HD to increase this risk by 1% using the upper 95% confidence limit of the β (0.025) is 2.9 ppm MeHg in maternal hair above and beyond the level to which the individuals in the reference group are exposed (median = 2.2 ppm). Alternatively, the background risk $R(0)$ of an abnormal score on two or more tests in the reference group for “total cases” is $6/144 = 4.2\%$ (see Table 3). The HD to increase this risk by 1% using the same parameter estimate as above is 9.5 ppm MeHg in maternal hair. We used the 95% upper confidence limit for our slope estimate; the point estimate indicated a slightly beneficial effect. It is apparent from these calculations that the estimates of a hazardous dose are directly dependent on the background risk and the additional risk that is deemed acceptable. That is, the HD increases with decreasing background risk and increasing acceptable additional risk.

The simple formulas described above should facilitate the use of complex epidemiological data in quantitative risk assessment.

The approach described here also has some limitations, most importantly the loss of statistical power by collapsing continuous outcome data into an ordinal variable. The relatively wide confidence intervals, in particular for the category specific odds ratios, demonstrate the lack of statistical precision. Nevertheless, analyzing ordinal variables in logistic regression is more efficient than analyzing binary variables generated from continuous data (Ananth and Kleinbaum 1997; Scott et al. 1997). It should be recognized, however, that the models used are not ‘biologically-based’ and have no relationship to specific biological events in child development. Similarly, we assumed that all endpoints are equally important given the scientific uncertainty about which endpoints are biologically most relevant with regard to the MeHg toxicity. Furthermore, we had to exclude children for whom data was missing on one or more developmental tests across the board, whereas in our original analyses children would be excluded in the analysis of some developmental endpoints but not others thereby maximizing our use of the recruited study population. Finally, our approach may be most useful when a large number of tests have been performed within functional domains. When only few tests within a domain are considered and the issue of multiple comparisons is not as important in the interpretation of findings, generalized linear models are a more appropriate choice.

In conclusion, we present an alternative characterization of the impact of prenatal mercury exposure on child development that could complement existing risk assessment strategies. Using this method of analysis we did not find an adverse association of prenatal MeHg exposure with child development in a population with a mean prenatal exposure of 6 ppm. We hope that our proposed methods are useful when the interpretation of study findings is likely to be complicated by multiple hypothesis tests, and in quantitative risk assessment of non-cancer health outcomes using epidemiological data.

Acknowledgments This research was supported by Grants 2R01-ES008442-05; R01-ES10219; R01-ES08442 and ES-01247 from the US National Institutes of Health; 1 UL1 RR024160-02 from the National Center for Research Resources; the Food and Drug Administration; US Department of Health and Human Services, and by the Ministry of Health, Republic of Seychelles.

References

- Ananth CV, Kleinbaum DG (1997) Regression models for ordinal responses: a review of methods and applications. *Int J Epidemiol* 26:1323–1333. doi:10.1093/ije/26.6.1323
- Axelrad DA, Bellinger DC, Ryan LM et al (2007) Dose-response relationship of prenatal mercury exposure and IQ: an integrative analysis of epidemiologic data. *Environ Health Perspect* 115:609–615
- Berry DA, Hochberg Y (1999) Bayesian perspectives on multiple comparisons. *J Stat Plan Inference* 82:215–227. doi:10.1016/S0378-3758(99)00044-0
- Budtz-Jorgensen E, Grandjean P, Keiding N et al (2000) Benchmark dose calculations of methylmercury-associated neurobehavioural deficits. *Toxicol Lett* 112/113:193–199. doi:10.1016/S0378-4274(99)00283-0
- Budtz-Jorgensen E, Keiding N, Grandjean P (2001) Benchmark dose calculation from epidemiological data. *Biometrics* 57:698–706. doi:10.1111/j.0006-341X.2001.00698.x
- Budtz-Jørgensen E, Keiding N, Grandjean P, Weihe P (2002) Estimation of health effects of prenatal methylmercury exposure using structural equation models. *Environ Health* 1(1):2
- Cernichiari E, Toribara TY, Liang L et al (1995) The biological monitoring of mercury in the Seychelles study. *Neurotoxicology* 16:613–628
- Clarkson TW (2002) The three modern faces of mercury. *Environ Health Perspect* 110(Suppl 1):11–23
- Cohen JT, Bellinger DC, Shaywitz BA (2005) A quantitative analysis of prenatal methyl mercury exposure and cognitive development. *Am J Prev Med* 29:353–365. doi:10.1016/j.amepre.2005.06.007
- Counter SA, Buchanan LH (2004) Mercury exposure in children: a review. *Toxicol Appl Pharmacol* 198:209–230. doi:10.1016/j.taap.2003.11.032
- Crump KS, Kjellstrom T, Shipp AM et al (1998) Influence of prenatal mercury exposure upon scholastic and psychological test performance: benchmark analysis of a New Zealand cohort. *Risk Anal* 18:701–713. doi:10.1023/B:RIAN.0000005917.52151.e6
- Crump KS, Van Landingham C, Shamlaye C et al (2000) Benchmark concentrations for methylmercury obtained from the Seychelles child development study. *Environ Health Perspect* 108:257–263. doi:10.2307/3454443
- Davidson PW, Myers GJ, Cox C et al (1998) Effects of prenatal and postnatal methylmercury exposure from fish consumption on neurodevelopment: outcomes at 66 months of age in the Seychelles child development study. *JAMA* 280:701–707. doi:10.1001/jama.280.8.701
- Efron B, Tibshirani R, Storey JD et al (2001) Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* 96:1151–1160. doi:10.1198/016214501753382129
- Gelman A, Tuerlinckx F (2000) Type S error rates for classical and Bayesian single and multiple comparison procedures. *Comput Stat* 15:373–390. doi:10.1007/s001800000040
- Glantz SA (2002) A primer of biostatistics. McGraw-Hill, New York
- Grandjean P, Weihe P, White RF et al (1997) Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury. *Neurotoxicol Teratol* 19:417–428. doi:10.1016/S0892-0362(97)00097-4
- McDowell MA, Dillon CF, Osterloh J et al (2004) Hair mercury levels in U.S. children and women of childbearing age: reference range data from NHANES 1999–2000. *Environ Health Perspect* 112:1165–1171
- Myers GJ, Davidson PW, Cox C et al (2003) Prenatal methylmercury exposure from ocean fish consumption in the Seychelles child development study. *Lancet* 361:1686–1692. doi:10.1016/S0140-6736(03)13371-5
- National Research Council (2000) Toxicological effects of methylmercury. National Academy Press, Washington, DC
- Perneger TV (1998) What’s wrong with Bonferroni adjustments. *BMJ* 316:1236–1238
- Rothman KJ (1986) Modern epidemiology. Little Brown, Boston

- Rothman KJ (1990) No adjustments are needed for multiple comparisons. *Epidemiology* 1:43–46
- Savitz DA, Olshan AF (1995) Multiple comparisons and related issues in the interpretation of epidemiologic data. *Am J Epidemiol* 142:904–908
- Savitz DA, Olshan AF (1998) Describing data requires no adjustment for multiple comparisons: a reply from Savitz and Olshan. *Am J Epidemiol* 147:813–814 discussion 815
- Scott SC, Goldberg MS, Mayo NE (1997) Statistical assessment of ordinal outcomes in comparative studies. *J Clin Epidemiol* 50:45–55. doi:[10.1016/S0895-4356\(96\)00312-5](https://doi.org/10.1016/S0895-4356(96)00312-5)
- Shamlaye C, Davidson PW, Myers GJ (2004) The Seychelles child development study: two decades of collaboration. *SMDJ Seychelles Med Dent J* 7:92–99
- Stokes ME, Davis CS, Koch GG (2000) Categorical data analysis using the SAS system. SAS Institute, Inc., Cary
- Thompson JR (1998) Invited commentary: Re: Multiple comparisons and related issues in the interpretation of epidemiologic data. *Am J Epidemiol* 147:801–806
- Thurston SW, Ruppert D, Davidson PW (2009) Bayesian models for multiple outcomes nested in domains. *Biometrics*
- van Wijngaarden E, Hertz-Picciotto I (2004) A simple approach to performing quantitative cancer risk assessment using published results from occupational epidemiology studies. *Sci Total Environ* 332:81–87. doi:[10.1016/j.scitotenv.2004.04.005](https://doi.org/10.1016/j.scitotenv.2004.04.005)
- van Wijngaarden E, Beck C, Shamlaye CF et al (2006) Benchmark concentrations for methyl mercury obtained from the 9-year follow-up of the Seychelles child development study. *Neurotoxicology* 27:702–709. doi:[10.1016/j.neuro.2006.05.016](https://doi.org/10.1016/j.neuro.2006.05.016)
- Veazie PJ (2006) When to combine hypotheses and adjust for multiple tests. *Health Serv Res* 41:804–818. doi:[10.1111/j.1475-6773.2006.00512.x](https://doi.org/10.1111/j.1475-6773.2006.00512.x)
- WHO (1990) Environmental health criteria 101 methylmercury. World Health Organization, Geneva