Estimation of the benchmark dose by structural equation models

ESBEN BUDTZ-JØRGENSEN

Department of Biostatistics, University of Copenhagen, Øster Farimagsgade 5, entrance B, PO Box 2099, DK-1014 Copenhagen K, Denmark and Institute of Public Health, University of Southern Denmark, Winslowparken 17, DK-5000 Odense C, Denmark

SUMMARY

While epidemiological data typically contain a multivariate response and often also multiple exposure parameters, current methods for safe dose calculations, including the widely used benchmark approach, rely on standard regression techniques. In practice, dose–response modeling and calculation of the exposure limit are often based on the seemingly most sensitive outcome. However, this procedure ignores other available data, is inefficient, and fails to account for multiple testing. Instead, risk assessment could be based on structural equation models, which can accommodate both a multivariate exposure and a multivariate response function. Furthermore, such models will allow for measurement error in the observed variables, which is a requirement for unbiased estimation of the benchmark dose. This methodology is illustrated with the data on neurobehavioral effects in children prenatally exposed to methylmercury, where results based on standard regression models cause an underestimation of the true risk.

Keywords: Environmental epidemiology; Measurement error; Multiple endpoints; Risk assessment.

1. INTRODUCTION

A key goal in statistical analyses used for quantitative risk assessment is to provide a scientific approach to setting safe exposure limits. Although experimental laboratory data may be used for this purpose, more complex epidemiological data on human health effects of environmental toxicant exposures are sometimes available. In recent years, the so-called benchmark approach has become the standard procedure for calculation of exposure limits (NAS, 2000). The benchmark dose (BMD) is defined as the dose that leads to a specific increased risk over the background (Crump, 1984). For a given data set, the BMD is estimated in a dose–response model, and statistical uncertainty is taken into account by basing the exposure limit on a lower confidence limit of the BMD, the so-called BMDL.

Given single exposure and response variables, the benchmark calculation is straightforward to perform. However, epidemiological data often contain more than one endpoint. A standard approach therefore involves selection of the outcome that appears most sensitive. The exposure limit is then based on the benchmark results for that outcome. Likewise, if more than one exposure parameter is available, a further choice needs to be made in this regard. This post hoc approach is clearly not attractive. Separate analyses of the outcomes are not efficient, and, because the multiple testing problem is ignored, the final
results may be criticized for being overly protective. In this regard, Crump (2002) noted that it would generally be a mistake to select the lowest BMDL, as this could favor results with a high degree of estimation uncertainty.

Recently, multivariate dose–response modeling has received increased attention in environmental epidemiology (Budtz-Jørgensen and others, 2002; Roy and others, 2003). For quantitative risk assessment in animal data, Regan and Catalano (2000) proposed a multivariate regression model allowing a separate dose–response function for each outcome component, while an unstructured covariance matrix described the dependence between outcomes obtained in the same subject. For each outcome \((Y_j, j = 1, \ldots, p)\), a threshold value \((t_j)\) defined an abnormal response, and the benchmark calculation focused on the combined risk that at least one outcome was abnormal (i.e. \(P(Y_1 > t_1 \lor \cdots \lor Y_p > t_p)\)). When applying this type of combined risk, each inclusion of an additional outcome will lead to an increase in the unexposed risk, which will typically lead to a decreased BMD (Budtz-Jørgensen and others, 2001). This consequence seems inappropriate if, for instance, all the responses reflect the same underlying variable. In that case, one would expect the BMD to remain stable, while the BMDL would be expected to increase as more outcomes are included and the underlying variable is better characterized. Furthermore, in epidemiological applications, the model proposed by Regan and Catalano will be too simple as it does not allow for measurement error in study variables. Budtz-Jørgensen and others (2004b) showed that failure to adjust for exposure error will lead to overestimation of the BMD (and the BMDL). In the following, we shall illustrate that the BMD calculation is also sensitive to response error. Thus, an unbiased analysis can only be obtained in models that allow for measurement error.

As a solution to these problems, we propose to conduct the risk assessment using structural equation models. In the measurement part of these models, observed variables are considered manifestations of the underlying latent variables. The structural part of the model then describes the relation between the latent variables and possibly a set of covariates. Thus, structural equation models allow analysis of the effect of a multivariate exposure on a multivariate response after accounting for measurement error (see Sanchez and others, 2005, for a recent review). These models therefore seem well suited for the BMD calculations, but this approach has apparently not been attempted before.

One of the controversial exposure limits relates to methylmercury from fish and seafood. In evaluating the evidence, the National Academy of Sciences (NAS, 2000) chose to emphasize the findings from a prospective epidemiological study in the Faroe Islands, where multiple exposure parameters and multiple plausible outcome variables were available. From toxicokinetic considerations, the mercury concentration in the cord blood was considered the most appropriate exposure variable, and other exposure information was ignored. Among a variety of outcomes, the expert committee chose to focus on the Boston Naming test (BNT) because it reflects complex cognitive functions and, at the same time, appeared to be the most sensitive outcome. Interindividual differences were taken into account by applying a standard uncertainty factor, by which the BMDL was divided to reach the recommended exposure limit. However, this default uncertainty factor does not take into account study-specific issues and therefore cannot counterbalance the deficiencies mentioned above.

2. THE BENCHMARK APPROACH TO SETTING SAFETY STANDARDS

This methodology was first developed for dichotomous (normal/abnormal) responses in experimental studies (Crump, 1984). The BMD is the dose that increases the risk of an abnormal response from \(P_0\) in unexposed subjects to \(P_0 + \text{BMR}\), where BMR is a constant to be specified, for example, by regulatory authorities. The statistical uncertainty is taken into account by basing the exposure limit calculation on the BMDL (1-sided lower 95% confidence limit of BMD).

In this paper, we focus on applications of the benchmark methodology in epidemiological data. Unlike experimental studies, observational studies usually do not include a control group completely free of
exposure; the responses show a continuous distribution, and they are influenced by confounders in addition to the exposure of interest. Budtz-Jørgensen and others (2001) described how these challenges can be overcome in regression models. Thus, let $Y(X, Z)$ denote the response of a subject with exposure $X$ and confounder values $Z = (Z_1, \ldots, Z_r)'$. We assume that the dependence of the response on the exposure and the confounders is given by a linear model

$$Y(X, Z) = \beta_0 + \beta_x g(X) + \beta_1 Z_1 + \cdots + \beta_r Z_r + \epsilon,$$  

(2.1)

where $\epsilon \sim N(0, \sigma^2)$, and the dose–response function $g$ is known and increases with $g(0) = 0$. Without loss of generality, we consider large responses to be disadvantageous. An adverse exposure effect therefore corresponds to a positive value of $\beta_x$.

In order to define the BMD for continuous outcomes, the level of an abnormal response $t_0$ must be specified. If $t_0$ is the same for all values of $Z$, then the unexposed risk $P\{Y(0, Z) > t_0\}$ will depend on the confounders, and so will the BMD, which is not practical. Instead, a covariate-dependent threshold is defined so that all unexposed subjects have the same prespecified risk $P_0$ (i.e. $P\{Y(0, Z) > t_0(Z)\} = P_0$ for all $Z$). In agreement with the original definition for dichotomous data, the BMD is then defined as the dose which increases the risk by BMR, that is, $P\{Y(BMD, Z) > t_0(Z)\} = P_0 + \text{BMR}$. Independent of the confounders, this dose is given by

$$\text{BMD} = g^{-1}(\Omega \sigma / \beta_x), \quad \text{if } \beta_x > 0,$$  

(2.2)

where $\Omega = \Phi^{-1}(1 - P_0) - \Phi^{-1}(1 - P_0 - \text{BMR})$. Often, $P_0$ is fixed at 5%, while the BMR is either 5% or 10% (NAS, 2000). Effect modification by a covariate $Z_j$ may be handled by adding an interaction term $\beta_{xj} g(X) Z_j$ to model (2.1). With this extension, subjects are not equally sensitive, and the BMD will depend on the level of the effect modifier, that is, $\text{BMD} = g^{-1}[\Omega \sigma / (\beta_x + \beta_{xj} z_j)], j = 1, \ldots, r$.

The estimate of the BMD in (2.2) is a monotone function of $\beta_x/\hat{\sigma}$, which follows a (scaled) noncentral $t$-distribution. Using a precise approximation to this distribution, Budtz-Jørgensen and others (2001) derived a lower confidence limit

$$\text{BMDL} = g^{-1}[\Omega \hat{\sigma}(1 - c_1)/[\hat{\beta}_x + u_{95} \hat{\text{SE}}(\hat{\beta}_x) \sqrt{1 + c_2}]], \quad \text{if } t = \hat{\beta}_x / \hat{\text{SE}}(\hat{\beta}_x) > -u_{95},$$  

(2.3)

where $c_1 = u_{95}^2/2\text{df}$, $c_2 = (t^2 - u_{95}^2)/2\text{df}$, $u_{95} \approx 1.645$ is the 95th percentile in the standard normal distribution, df is the number of degrees of freedom, and $\hat{\text{SE}}(\hat{\beta}_x)$ is the estimated standard error of $\hat{\beta}_x$. Crump (2002) recommended confidence limit calculations based on the (profile) likelihood function and showed that this approach yielded results that were similar to (2.3).

3. STANDARD APPLICATION OF THE BENCHMARK APPROACH IN EPIDEMIOLOGICAL DATA

In the Faroe Islands, the population is exposed to increased levels of methylmercury mainly through consumption of contaminated pilot whale meat. During 1986–1987, a birth cohort of 1022 Faroese children was established for prospective studies to examine possible adverse effects of prenatal exposure to methylmercury. The intrauterine methylmercury exposure was determined by the analysis of umbilical cord blood and maternal hair for mercury. At the age of 7 years, the cohort members participated in a thorough neuropsychological examination consisting of a number of tests assessing different domains of brain function. Standard multiple regression analysis showed strong mercury effects especially for the BNT (Grandjean and others, 1997, 1999). A NAS (2000) committee identified the Faroese study as the critical study for calculating an exposure limit for methylmercury. This expert committee also recommended that this limit should be derived based on the benchmark approach.

Table 1 shows benchmark results for each of the verbal outcomes in the Faroese data assuming a linear $\{g(x) = x\}$ or a logarithmic $\{g(x) = \log(x+1)\}$ dose–response function, respectively. Furthermore,
Table 1. BMD and BMDL calculated for cord blood concentration (µg/l) and the maternal hair mercury concentration (µg/g) assuming a linear and a logarithmic dose–response regression model for the individual outcome variables. \( P_0 \) and BMR were fixed at 5%.

<table>
<thead>
<tr>
<th>Response</th>
<th>Cord blood Hg</th>
<th>Maternal hair Hg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear</td>
<td>Logarithmic</td>
</tr>
<tr>
<td></td>
<td>BMD</td>
<td>BMDL</td>
</tr>
<tr>
<td>Wechsler Intelligence Scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digit spans (DS)</td>
<td>222</td>
<td>99.9</td>
</tr>
<tr>
<td>BNT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No cues (BNT1)</td>
<td>102</td>
<td>64.9</td>
</tr>
<tr>
<td>With cues (BNT2)</td>
<td>84.9</td>
<td>57.6</td>
</tr>
<tr>
<td>California Verbal Learning Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning (CVLT1)</td>
<td>404</td>
<td>125</td>
</tr>
<tr>
<td>Short-term reproduction (CVLT2)</td>
<td>219</td>
<td>99.1</td>
</tr>
<tr>
<td>Long-term reproduction (CVLT3)</td>
<td>250</td>
<td>104</td>
</tr>
<tr>
<td>Recognition (CVLT4)</td>
<td>228</td>
<td>99.7</td>
</tr>
</tbody>
</table>

\( P_0 \) and the BMR were fixed at 5%, so that exposure to the BMD would double the unexposed risk (i.e. from 5% to 10%). The results are seen to be heavily dependent on the dose–response model with the logarithmic model yielding the lower BMDLs. The NAS committee did not find this model to be biologically plausible and instead recommended the linear model. Even when the calculations were restricted to this model, the BMDLs varied strongly between the Faroese neuropsychological test scores. The committee decided to base the final calculation on the outcome with the strongest mercury effect (i.e. the cued BNT score). This procedure yielded a BMDL of 58 µg/l in cord blood, which formed the basis for the final limit on the daily human mercury intake (NAS, 2000).

4. LIMITATIONS OF STANDARD BENCHMARK CALCULATIONS

Standard regression approaches will produce a BMDL for each exposure–outcome pair. The NAS committee’s decision to base the final calculation on the outcome with the most significant mercury effect disregards multiple testing problems and may therefore produce exposure limits that are overly protective. On the other hand, the regression approach estimates the mercury effect on each outcome separately. This method is not statistically efficient, as information is not pooled across different regressions. Application of a less efficient method will produce a lower BMDL, so all in all it is not clear whether this naive procedure is conservative or not. As will be illustrated below, the situation is even more complicated when the exposure or response variables are affected by the measurement error.

A main advantage of the benchmark approach as compared to older methods is that the uncertainty arising from limitations in the sample size is taken into account. A smaller study will tend to yield lower and more protective exposure limits. However, a limited sample size is not the only source of uncertainty affecting epidemiological studies. Typically, study variables cannot be measured without error. It is well known that nondifferential measurement error in the exposure variable will lead to underestimation of exposure effects in regression analysis (Carroll and others, 1995). In agreement with this result, Budtz-Jørgensen and others (2004b) showed that failure to take exposure measurement error into account will lead to BMDs and BMDLs that are biased toward higher levels. This bias increases as a function of the measurement error variance but decreases as a function of the true exposure variation. Budtz-Jørgensen and others (2004b) derived a flexible adjustment method in nonlinear regression models and estimated
that, due to uncertainty in the assessment of prenatal mercury exposure, the BMDL recommended by the NAS committee was about 22% too high.

Epidemiological outcome variables may also be affected by measurement error, but this is usually not emphasized as regression coefficients are not biased. However, benchmark calculations depend on a reliable estimate of the true response variation (Gaylor and Slikker, 2004) and are therefore also sensitive to measurement uncertainty in the response. This effect can be illustrated by assuming an additive error in the relationship between the true response ($\eta$) and the measured response $Y$, that is $Y(X, Z) = \eta(X, Z) + \zeta$, where $\zeta$ is a random error with a mean of zero and variance $\psi^2$. As a result of this added uncertainty, the conditional variance in the response becomes $\sigma^2 + \psi^2$, where $\sigma^2$ is the variance in $\eta$ given $X$ and $Z$. The estimated cutoff ($t_0$) defining an adverse response will satisfy $P\{Y(0, Z) > t_0\} = 5\%$ and will therefore be overestimated compared to the value of the error-free situation defined by $P\{\eta(0, Z) > t_0\} = 5\%$. Although the additional variation will not bias the dose–response function, the BMD will be overestimated.

Fig. 1. Benchmark calculation based on the true response (top graph) and an error-prone measure (bottom graph). The BMD increases the risk of a response above the threshold from $P_0$ to $P_0 + \text{BMR}$. Measurement error increases the response variation. As a result, the BMD and the threshold defining an adverse response ($t_0$) are overestimated.
because the same increase in the mean will lead to a weaker increase in the probability above the cutoff. Thus, a higher dose will be needed in order to change the risk from \( P_0 \) to \( P_0 + \text{BMR} \) (Figure 1). While the true BMD is \( g^{-1}(\Omega/\beta) \), the naive analysis estimates \( g^{-1}(\Omega\sqrt{\sigma^2 + \psi^2/\beta}) \). The consequences for the lower confidence limit, the BMDL, are less obvious because the upward bias introduced in the BMD estimate is accompanied by an increased estimation uncertainty, which will push the BMDL toward zero. However, as noted by Budtz-Jørgensen and others (2001), the BMDL increases as a function of the residual variance in the response. So the total effect is that a more uncertain response will lead to higher BMDLs.

5. A STRUCTURAL EQUATION MODEL

In epidemiological risk assessment, typically data collected for the \( i \)th subject consist of both a multivariate response \( Y_i = (Y_{i1}, \ldots, Y_{ip})' \) and a multivariate exposure \( X_i = (X_{i1}, \ldots, X_{iq})' \) as well as a set of covariates \( Z_i = (Z_{i1}, \ldots, Z_{ir})' \), \( i = 1, \ldots, n \). We propose to do the benchmark calculation based on a structural equation model for the conditional distribution of \((X_i, Y_i)\) given \( Z_i \). Structural equation models generally consist of a measurement part and a structural part (Muthén, 1984; Bollen, 1989). In the measurement part, observed variables are viewed as manifestations of a limited number of latent variables, which are then assumed to be causally related in the structural part of the model. Thus, the measurement model is

\[
\begin{align*}
    g(X_{i1}) &= \mu_1 + \lambda_{11} g(\eta_{i1}) + \epsilon_{i1} \\
    \vdots & \quad \vdots \\
    g(X_{iq}) &= \mu_q + \lambda_{q1} g(\eta_{i1}) + \epsilon_{iq} \\
    Y_{i1} &= \nu_{q1} + \lambda_{q12} \cdot \eta_{i2} + \epsilon_{i(q+1)} \\
    \vdots & \quad \vdots \\
    Y_{ip} &= \nu_{q+p} + \lambda_{q+p2} \cdot \eta_{i2} + \epsilon_{i(q+p)}.
\end{align*}
\]

(5.1)

In matrix notation, \((g(X_i)', Y_i')' = v + \Lambda (g(\eta_{i1}), \eta_{i2})' + \epsilon_i\), where \( v_{(q+p) \times 1} \) and \( \Lambda_{(q+p) \times 2} \) are parameter matrices. Thus, observed variables \( X_i \) and \( Y_i \) are indicators of an underlying true exposure \((\eta_{i1})\) and an underlying response function \((\eta_{i2})\), respectively. The error term \( \epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{i(q+p)})' \) follows a normal distribution with \( E(\epsilon_i) = 0 \) and \( \text{cov}(\epsilon_i) = \Psi \). The regression coefficients \((\lambda_{ij})\) are known as factor loadings. These allow observed variables to have different scales.

The structural part of the model assumes a linear effect of the true transformed exposure on the latent response, while covariates are allowed to affect both latent variables

\[
\begin{align*}
    g(\eta_{i1}) &= a_1 + \gamma_{11} Z_1 + \cdots + \gamma_{1r} Z_r + \zeta_1, \\
    \eta_{i2} &= a_2 + \beta_{21} g(\eta_{i1}) + \gamma_{21} Z_1 + \cdots + \gamma_{2r} Z_r + \zeta_2.
\end{align*}
\]

(5.2)

In matrix notation, \((g(\eta_{i1}), \eta_{i2})' = a + B g(\eta_{i1}) + \Gamma Z_i + \zeta_i\), where \( a_{2 \times 1}, B_{2 \times 2}, \) and \( \Gamma_{2 \times r} \) are parameter matrices. All elements in \( B \) are zero except for the element \((2, 1)\), which is the exposure effect \( \beta_{21} \). The dose–response function \( g \) is known and increases with \( g(0) = 0 \). The residual \( \zeta_i = (\zeta_{i1}, \zeta_{i2})' \) is independent of \( \epsilon_i \) and follows a normal distribution with \( E(\zeta_i) = 0 \) and \( \text{cov}(\zeta_i) = \Psi = \text{diag}(\psi_1^2, \psi_2^2) \).

Figure 2 shows a path diagram illustrating the proposed model. In the path diagrams, observed variables are in boxes, while latent variables are in circles or ovals. Thus, after transformation with \( g \), exposure variables are assumed to be linearly related to the underlying (transformed) true exposure and a random measurement error. Similarly, outcome variables are considered indicators of a latent response function, which is assumed to depend on the latent exposure. Thus, the dose–response relationship is modeled using only one effect parameter, and the analysis allows for measurement error both in the exposure and in the outcome variables. As can be seen from (5.1) and (5.2), if a given dose–response function \( g \) is selected in the structural part of the model, then the linear structure of the model implies that the measurement part
Estimation of the BMD

Fig. 2. Path diagram illustrating a structural equation model for benchmark analysis. Observed exposures \((X_1, \ldots, X_q)\) and outcomes \((Y_1, \ldots, Y_p)\) are assumed to depend on underlying latent variables \((\eta_1, \eta_2)\) and a random measurement error \((\epsilon_1, \ldots, \epsilon_{q+p})\). After transformation with \(g\), the latent exposure \((\eta_1)\) has a linear effect on the latent response \((\eta_2)\). Covariates \((Z_1, \ldots, Z_r)\) can be related to both the latent functions.

holds after transformation with \(g\). Thus, in this framework the fit of the measurement model must be taken into consideration when selecting the dose–response function.

5.1 Inference and identifiability

The model parameters are given by \(\theta = (\nu, \Lambda, \Upsilon, \alpha, B, \Gamma, \Psi)\). The distribution of \((X_i, Y_i)\) given \(Z_i\) is 
\[ N_{q+p}(\mu(\theta) + \Pi(\theta)z_i, \Sigma(\theta)), \]
where \(\mu(\theta) = \nu + \Lambda(I - B)^{-1}\alpha, \Pi(\theta) = \Lambda(I - B)^{-1}\Gamma, \) and \(\Sigma(\theta) = \Lambda(I - B)^{-1}\Psi(I - B)^{-1}\Lambda^T + \Upsilon\). Therefore, the likelihood function is
\[ L(\theta) \propto \prod_{i=1}^{n} \exp\left\{-[v_i - \mu(\theta) - \Pi(\theta)z_i]^T \Sigma(\theta)^{-1}[v_i - \mu(\theta) - \Pi(\theta)z_i]/2\right\}/\sqrt{\left|\Sigma(\theta)\right|}, \]  
(5.3)

where \(v_i = (x_i^T, y_i^T)^T\) and \(\left|\cdot\right|\) denotes the determinant. Maximum likelihood estimates can be obtained using standard numerical maximization techniques such as Newton–Raphson or Fisher scoring (Bollen, 1989). Standard errors are estimated from the inverse of the Fisher information, and the significance of individual parameters can be assessed using Wald or likelihood ratio testing. When \(v_i\) is incomplete and values can be assumed to be missing at random (Little and Rubin, 2002), the contribution to the likelihood of the \(i\)th subject is determined by integrating the \(i\)th factor in (5.3) with respect to the missing variables. This is achieved simply by deleting rows and columns corresponding to the missing variables in \(\mu(\theta), \Pi(\theta), \) and \(\Sigma(\theta)\). The resulting function can then be maximized using the expectation maximization algorithm (Little and Rubin, 2002). Model fit is commonly assessed in a likelihood ratio test of the proposed model against a more flexible model with no structure on \(\mu, \Pi, \) and \(\Sigma\).

Structural equation models rely on unobservable variables and are therefore subject to identifiability problems. To avoid this, a scale and a zero point must be selected for each latent variable. By fixing \(\lambda_{11} = 1\), the latent exposure \(\eta_1\) is expressed on the scale of \(X_1\), in the sense that a 1-unit increase in \(g(\eta_1)\) (on average) leads to a 1-unit increase in \(g(X_1)\). Similarly, the zero point of \(g(\eta_1)\) may be defined by fixing \(v_1 = 0\). Even after such restrictions, the model may not be identified. The requirement is that the mapping \(\theta \rightarrow \{\mu(\theta), \Pi(\theta), \Sigma(\theta)\}\) must be one-to-one. In general, this may be hard to prove (Bollen, 1989), but if the observed indicators are assumed to have independent measurement errors (\(\Upsilon\) is diagonal matrix), then the model above is identified if 2 or more indicators are available for each of the latent variables.
Here, BMD analysis is conducted in a structural equation model with reference indicators $X_1$ and $Y_1$, that is $\lambda_{11} = \lambda_{(g+1)1} = 1$ and $v_1 = v_{(g+1)} = 0$. The structural part of the model describing the relation between true exposure and the underlying outcome function is $\eta_2 = a_2 + \beta_{21} g(\eta_1) + \gamma_{21} Z_1 + \cdots + \gamma_{2r} Z_r + \zeta_2$. Although this relationship involves latent variables, it has the same form as the regression model (6.1), and therefore the benchmark methodology of Section 2 can be directly applied. Thus, exposure to the BMD increases the risk $P(\eta_2 > t_0(Z))$ from $P_0$ in an unexposed subject to $P_0 + BMR$. In the scale of the reference exposure indicator ($X_1$), the BMD expression is similar to the regression result, that is

$$BMD = g^{-1}(\Omega \psi_2 / \beta_{21}),$$

where $\psi_2 = \sqrt{\text{var}(\zeta_2)}$. A consistent estimator of the BMD is obtained by replacing the unknown parameters in (6.1) by maximum likelihood estimates.

The BMD may also be expressed on a scale different from that of $X_1$. Had $X_j$ been selected as the reference indicator, the exposure effect and the intercept in the structural part of the model (5.2) would have been, respectively, $\beta_{21}/\lambda_{jj}$ and $a_2 - \beta_{21} v_1 / \lambda_{jj}$. In this parameterization, the BMD is $g^{-1}(\Omega \psi_2 / \beta_{21})$. It should be noted that the benchmark calculation does not depend on the reference indicator of the latent response. A change in the reference outcome variable would increase the exposure effect parameter and the residual standard deviation in the latent response by the same factor and therefore does not change the BMD.

The BMDL may be calculated based on the $\delta$-method. Because $\hat{\beta}_{21} / \hat{\psi}_2$ is often approximately normal, an upper confidence limit was first calculated for $\beta_{21} / \psi_2$, and then transformed into a BMDL. This process leads to

$$\text{BMDL} = g^{-1}\left[\Omega \hat{\psi}_2 / (\hat{\beta}_{21} + u_05 \sqrt{\text{var}(\hat{\beta}_{21}) - \text{cov}(\hat{\psi}_2, \hat{\beta}_{21}) \hat{\beta}_{21} / \hat{\psi}_2^2 + \text{var}(\hat{\psi}_2) \hat{\beta}_{21}^2 / 4 \hat{\psi}_2^4)}\right],$$

where $\text{var}(\hat{\beta}_{21})$, $\text{var}(\hat{\psi}_2^2)$, $\text{cov}(\hat{\psi}_2, \psi_2^2)$ are estimates of the variance in $\hat{\beta}_{21}$ and $\hat{\psi}_2^2$, and the covariance between them, respectively. The BMDL may be expressed in the scale of exposure variable $X_j$ by replacing $(\hat{\beta}_{21}, \text{var}(\hat{\beta}_{21}), \text{cov}(\hat{\beta}_{21}, \hat{\psi}_2^2))$ by $(\hat{\beta}_{21} / \lambda_{jj}, \text{var}(\hat{\beta}_{21} / \lambda_{jj}), \text{cov}(\hat{\beta}_{21} / \lambda_{jj}, \hat{\psi}_2^2 / \hat{\lambda}_{jj}))$ in (6.2).

While the $\delta$-method offers a computationally simple solution, confidence limits based on the likelihood function are generally recommended. These may be calculated after a reparameterization of the model. Instead of fixing the factor loading of $Y_1$ at one, the scale of $\eta_2$ is defined by fixing the residual variance $\psi_2^2 = 1$. With this parameterization, the BMD depends on only one parameter, that is $BMD = g^{-1}(\Omega / \beta_{21})$. Letting $\hat{\theta} = \theta \setminus \{\beta_{21}\}$, the profile likelihood function for $\beta_{21}$ is given by $L_p(\beta_{21}) = \max_{\hat{\theta}} L(\hat{\theta}, \beta_{21})$, where $L$ is given by (5.3) possibly modified to allow for incomplete data. An upper confidence limit for $\beta_{21}$ is determined by

$$\beta_{21}^{up} = \max[\beta_{21} : 2\log(L(\hat{\theta})) - \log(L_p(\beta_{21}))] \leq \chi_1^2(95),$$

where $\chi_1^2(95) \approx 3.841$ is the 95th percentile in the $\chi^2$-distribution with 1 degree of freedom. A lower 95% confidence limit for BMD may then be obtained as $\text{BMDL} = g^{-1}(\Omega / \beta_{21}^{up})$. The value of $\beta_{21}^{up}$ may be computed using an algorithm of Neale and Miller (1997), which has been incorporated into the free software package Mx (Neale and others, 2003).

Finally, the BMDL may also be calculated by sampling a large number of bootstrap data sets with replacement from the original data. In each data set, the BMD is estimated based on the expression (6.1), and the BMDL can then be calculated as the 5th percentile in the empirical distribution of the estimates. Standard inference in structural equation models depends on multivariate normality assumptions. While the maximum likelihood estimator is consistent as long as the first 2 moments of the data have been correctly specified, estimation precision is often overestimated in nonnormal data (Arminger and
Schoenberg, 1989). Thus, by basing the BMDL calculation on the nonparametric bootstrap, the analysis is less likely to be sensitive to deviations from multivariate normality.

6.1 Simulation study

A small simulation study was conducted to compare the statistical properties of the likelihood-based BMDL to that of the $\delta$-method. Data were simulated using the model of Figure 2 with 3 indicators for both $\eta_1$ and $\eta_2$. For simplicity, indicators had independent measurement errors, no covariates were included, and $g(x) = x$. The exposure effect was varied so that the standardized effect $\beta_{12}\sqrt{\text{var}(\eta_1)/\text{var}(\eta_2)}$ was 0.05, 0.25, or 0.5. To study the effects of measurement error, precisions were fixed, so that the ratio of true variation to total variation in the exposures, $R^2_x = \lambda^2_1\text{var}(\eta_1)/\text{var}(X_j)$, $j = 1, 2, 3$, and in the outcomes, $R^2_y = \lambda^2_{(3+k)2}\text{var}(\eta_1)/\text{var}(Y_k)$, $k = 1, 2, 3$, was either 0.50 or 0.90. The sample size ($n$) was 100, but for a standardized exposure effect of 0.5, sample sizes of 500 and 1000 were also considered.

BMDLs were calculated using the likelihood function, the $\delta$-method, and a naive approach similar to that of the NAS in the mercury example. First, the strongest association between exposure and response variables were identified based on $t$-tests, then, from this relationship, a regression-based BMDL was calculated using (2.3). In each data set, the BMDLs were compared to the known BMD, and the coverage probabilities were determined using 2000 independently generated data sets for each configuration of the parameters.

Table 2 shows that, even with a relatively limited sample size, the coverage probability of the BMDL is close to the nominal value of 95% for both the profile likelihood and the $\delta$-method. While the results of the likelihood-based method are very stable, for large effect sizes and high degrees of imprecision, there is a tendency for the coverage probability of the $\delta$-method to decrease. In this regard, exposure imprecision seems more important than outcome error. However, as the sample sizes increase, the coverage probability approaches 95%. In environmental epidemiology, standardized effect sizes above 20% are rare, so in general the $\delta$-method would seem to be sufficient. A strong variation is seen in coverage probability of the naive method. While the coverage probability is very low at high exposure effects, the results indicate that this method is overly protective at small effect sizes. As expected, coverage probabilities decrease as a function of the imprecision in observed variables. Results for this approach are also more sensitive to exposure error. It should be noted that the results for the naive approach are heavily dependent on sample size. Thus, for a standardized effect size of 0.5 and a sample size of 1000, the coverage probability was below 1% irrespective of the imprecision (data not shown).

Table 2. Coverage probability (in %) of BMDLs calculated using the profile likelihood (pl), the $\delta$-method ($\delta$), and the naive method based on the strongest exposure–response association (naive). Results are given for different choices of the standardized exposure effect (effect) and different degrees of precision in exposure ($R^2_x$) and response variables ($R^2_y$). The sample size ($n$) was 100 unless otherwise stated. The number of data sets generated was 2000, yielding a standard error of 0.487% in the coverage probability

<table>
<thead>
<tr>
<th>$R^2_x$ (%)</th>
<th>$R^2_y$ (%)</th>
<th>Effect = 5%</th>
<th>Effect = 25%</th>
<th>Effect = 50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>pl</td>
<td>$\delta$</td>
<td>naive</td>
<td>pl</td>
<td>$\delta$</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>$n = 1000$</td>
<td>$n = 500$</td>
<td>$n = 1000$</td>
<td>$n = 500$</td>
</tr>
<tr>
<td>50, 50</td>
<td>94.30</td>
<td>94.40</td>
<td>99.95</td>
<td>94.45</td>
</tr>
<tr>
<td>50, 90</td>
<td>94.30</td>
<td>93.90</td>
<td>99.30</td>
<td>95.00</td>
</tr>
<tr>
<td>90, 50</td>
<td>94.50</td>
<td>94.55</td>
<td>99.75</td>
<td>96.20</td>
</tr>
<tr>
<td>90, 90</td>
<td>94.45</td>
<td>94.40</td>
<td>98.90</td>
<td>94.20</td>
</tr>
</tbody>
</table>
7. Application to the Faroese Data

We consider an extension of the structural equation model developed by Budtz-Jørgensen and others (2002, 2003) for estimating the effect of prenatal mercury exposure in the Faroese children. Test scores on the BNT, the Wechsler Intelligence Scale Digit Spans (DS), and the California Verbal Learning Test (CVLT) were considered manifestations of a latent verbal function. An acceptable model fit was obtained only after the inclusion of correlations between measurement errors in the 2 BNT scores and the 4 CVLT scores, respectively. This type of association is known as local dependence and was modeled by allowing nonzero off-diagonal elements in $\Upsilon$.

On the exposure side, mercury concentrations in cord blood and maternal hair were assumed to be linearly related to the underlying true exposure after transformation with the function $g$. Budtz-Jørgensen and others (2002, 2003) used $g(x) = \log(x)$ which does not satisfy $g(0) = 0$. The benchmark analysis was therefore based on a modified logarithmic model $g(x) = \log(x + 1)$. The NAS committee did not find logarithmic dose–response functions plausible, and the structural equation calculations were therefore conducted also for the linear model $g(x) = x$. The logarithmic model had a better fit mainly because it provided a superior measurement model for the exposure variables. Thus, the association between mercury concentrations in cord blood and maternal hair becomes approximately linear with a uniform scatter after a logarithmic transformation (Budtz-Jørgensen and others, 2004a). As for the relation between exposures and outcomes, scatter plots and standard regression techniques failed to show significant differences in the model fit between a linear and a logarithmic dose–response function. Finally, potential confounders were included in the model as covariates that were associated both with the latent exposure and with the latent response.

7.1 Results

Table 3 shows estimated mercury effects in the linear and the logarithmic model both before and after the inclusion of incomplete cases in a missing data analysis. The interpretation of the effect parameter ($\beta_{21}$) depends on the choice of reference indicators. Here, we have chosen to express the true exposure on the scale of the cord blood concentrations, while the verbal function is on the scale of the cued Boston Naming test (BNT2). Thus, for the logarithmic model, $\beta_{21}$ gives the expected change in the BNT2 score for each 10-fold increase in true cord-blood concentration plus one. For the linear model, $\beta_{21}$ is the expected change in BNT2 caused by an exposure increase of 10 $\mu$g/l. Standard errors were estimated both based on the likelihood function and from analyzing 1000 bootstrap resamples of the data. Irrespective of the standard error calculation, the mercury effect is highly significant in both models. For the linear model, where the fit of the measurement model for the exposures was not perfect, the standard errors of the bootstrap were higher than the model-based values. For the logarithmic model, a closer agreement was seen between the 2 sets of BMDL values.

Benchmark results of the structural equation models are given in Table 4. A nice agreement is seen between the likelihood and the $\delta$-method BMDL with the former providing slightly higher results. In agreement with the standard errors of Table 3, the BMDLs of the bootstrap are somewhat lower for the linear model, while the bootstrap provides results which are close to the other methods in the logarithmic model. As may have been expected based on the regression results, the BMDLs of the logarithmic model are clearly lower than those obtained in the linear model. More interestingly, the integrated structural equation result is lower than each of the individual regression results. For the linear model, the BMDL is about 40 $\mu$g/l, while the lowest individual BMDL is around 60 $\mu$g/l (Table 1). Thus, in this case, the standard procedure of selecting the BMDL of the most sensitive response is not protective. In the naïve BMDL calculation, the positive bias induced by ignoring measurement error outweighed the negative bias caused by not accounting for multiple testing.
Table 3. Mercury effects obtained in structural equation models before and after the inclusion of incomplete cases and with a linear and a logarithmic dose–response function, respectively. Reference indicators were (transformed) cord blood concentration for latent exposure and BNT cued for latent verbal function. The linear model estimates the effect for an exposure increase of 10 $\mu$g/l, while the logarithmic model considers the effect a 10-fold increase in true concentration $+1$. Standard errors (SEs) were estimated based on the model assumptions and using the nonparametric bootstrap.

<table>
<thead>
<tr>
<th>Dose–response model</th>
<th>Complete cases</th>
<th>Missing data analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$</td>
<td>model-SE</td>
</tr>
<tr>
<td>Linear</td>
<td>0.235</td>
<td>0.0703</td>
</tr>
<tr>
<td>Logarithmic</td>
<td>1.96</td>
<td>0.549</td>
</tr>
</tbody>
</table>

Table 4. Benchmark results obtained in structural equation models assuming either a linear or a logarithmic relationship between the true exposure concentration and a latent verbal function. BMDLs were calculated using the profile likelihood (pl), the $\delta$-method ($\delta$), and the bootstrap with complete cases only and after the inclusion of incomplete cases. $P_0$ and BMR both were fixed at 5%. Results are expressed in the scale of the cord blood concentration ($\mu$g/l) and the maternal hair mercury concentration ($\mu$g/g).

<table>
<thead>
<tr>
<th>Cord blood Hg</th>
<th>Maternal hair Hg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BMD</td>
</tr>
<tr>
<td>Linear model</td>
<td></td>
</tr>
<tr>
<td>Complete cases</td>
<td>58.9</td>
</tr>
<tr>
<td>MAR analysis</td>
<td>64.8</td>
</tr>
<tr>
<td>Logarithmic model</td>
<td></td>
</tr>
<tr>
<td>Complete cases</td>
<td>4.08</td>
</tr>
<tr>
<td>MAR analysis</td>
<td>4.48</td>
</tr>
</tbody>
</table>

Inclusion of incomplete cases had little effect in this study. As expected, BMDLs generally increased slightly as a result of including more information. However, it should be noticed that also the BMDs became slightly higher, so the BMDL increase is not a result of improved estimation precision only.

8. SUMMARY AND DISCUSSION

This paper proposes to calculate the BMD in structural equation models. In these models, observed variables are viewed as representations of a limited number of causally related latent variables. This structure allows information from a multivariate exposure and a multivariate response to be incorporated in a joint dose–response analysis. In this way, power may be gained, and measurement error in study variables can be taken into account. These properties are important for risk assessment because the benchmark calculation depends on the study power, and naive regression results become biased toward less protective levels if the exposure and the response are affected by measurement error. This problem was clearly illustrated in the Faroese data set, where the integrated structural equation result was lower than each of the individual BMDLs obtained using the regression analysis. Thus, the simple approach of selecting the BMDL of the most sensitive outcome may not be protective. In epidemiological data, which are always affected by measurement error, a structural equation analysis will therefore contribute important information to standard-setting efforts.
The proposed structural equation method may seem difficult to comprehend, given that safety limits will be calculated from unobservable latent variables. On the other hand, the dependence of benchmark results on imprecision in exposure and outcome variables makes interpretation of the unadjusted values ambiguous. Thus, latent variables would be advantageous in epidemiological applications. The structural equation model estimates the degree of uncertainty based on an assumption that some observed variables are related to a common latent variable. The relevance of the resulting BMD depends on the appropriateness of this model. Therefore, the grouping of manifest variables into latent variables should be based on \textit{a priori} biological knowledge, rather than \textit{post hoc} data-driven processes such as exploratory factor analysis (Jolliffe, 1998). For example, in the Faroese data, neurobehavioral outcome variables were grouped into major nervous system functions determined by the neuropsychologists involved in the project. The excellent fit of the proposed model suggests that the BMDs derived are appropriate for verbal functioning in children. Despite this, skeptical risk assessors may prefer standard regression techniques for the main safety limit calculation. Still, a structural equation analysis would provide an important supplement for illustration of the sensitivity to measurement errors.

In the structural equation approach, the benchmark calculation focuses on the probability that the underlying latent response is abnormal (i.e. \(P(\eta > t_0)\)). For experimental animal data, Regan and Catalano instead considered the combined risk that at least one of the observed indicators was abnormal (i.e. \(P(Y_1 > t_1 \lor \cdots \lor Y_p > t_p)\)). The latter probability is not appropriate for risk assessment when the observed variables are imprecise reflections of the same underlying variable, because the risk would increase with the number of indicators and with the degree of imprecision associated with these indicators. In contrast, the risk concept of the structural equation model does not depend on the quality of the instruments used for measuring the response or the exposure. However, for structural equation models with more than one latent response, a possible way to extend this methodology would be to base the benchmark calculation on the combined risk that (at least) one of the underlying variables had an abnormal level. For instance, Budtz-Jørgensen and others (2002) developed a structural equation model including indicators of both verbal and motor function in the Faroese data. In this model, it would be appropriate to base the risk assessment on the probability that one of the latent functions is abnormal.

Identification of safe exposure levels is an important extension of dose–response modeling from environmental epidemiology. The statistical methodology has improved somewhat in recent years, but current methods are not perfect. This paper has focused on the benchmark approach, which is one of the most widely used methods for setting standards. This method is clearly superior to older methods (Crump, 1984), but, in epidemiological applications, it is an important weakness that measurement error in the response or in the exposure will bias results toward higher and less protective exposure levels. Thus, a less precise study will lead to a higher exposure limit. This property is counter to the precautionary principle, as discussed by Barnett and O’Hagan (1997, p. 40), implying that weaker knowledge should lead to more stringent standards. As illustrated in the present study, this problem can be handled using structural equation modeling, but it may also be useful to develop new methods which are robust to measurement error. One possibility may be to base the calculations on hockey stick models, which assume no effect of the exposure below an unknown threshold. In the presence of exposure imprecision, the threshold is typically underestimated (Kuchenhoff and Carroll, 1997). Thus, contrary to the benchmark approach, the naive analysis that ignores the measurement error will produce lower hockey stick thresholds and thereby more protective standards. However, this method also suffers from the fact that exposure limits increase when the response is less precise (Keiding and Budtz-Jørgensen, 2003).

\textbf{ACKNOWLEDGMENTS}

I am grateful to Dr Pal Weihe for allowing me to use the data from the Faroese cohort study and to Drs Philippe Grandjean and Niels Keiding for inspiring discussions. I thank the editor and the anonymous
reviewers for useful advice. This study was supported by grants from the National Institute of Environmental Health Sciences (ES09797), the United States Environmental Protection Agency (9W-0262-NAEX), the European Commission (Environment Research Programme), and the Danish Medical Research Council. The contents of this paper are solely the responsibility of the author and do not necessarily represent the official views of the National Institute of Environmental Health Sciences, National Institutes of Health, or any other funding agency. Conflict of Interest: None declared.

REFERENCES


[Received October 11, 2005; first revision June 23, 2006; second revision September 12, 2006; third revision October 13, 2006; accepted for publication October 20, 2006]