Confounder selection in environmental epidemiology: Assessment of health effects of prenatal mercury exposure

Budtz-Jørgensen, Esben; Keiding, Niels; Grandjean, Philippe; Weihe, Pál

Published in:
Annals of Epidemiology

DOI:
10.1016/j.annepidem.2006.05.007

Publication date:
2007

Document version:
Accepted manuscript

Citation for pulished version (APA):
Budtz-Jørgensen, E., Keiding, N., Grandjean, P., & Weihe, P. (2007). Confounder selection in environmental epidemiology: Assessment of health effects of prenatal mercury exposure. *Annals of Epidemiology*, *17*(1), 27-35. https://doi.org/10.1016/j.annepidem.2006.05.007

Confounder Selection in Environmental Epidemiology: Assessment of Health Effects of Prenatal Mercury Exposure

PURPOSE: To compare different approaches to identification of confounders needed for analyzing observational data. While standard analysis is usually conducted as if the confounders were known *a priori* selection uncertainty must also be taken into account.

METHODS: Confounders were selected using backward elimination, the change in estimate method (CIE), Akaike's information criterion (AIC), the Bayesian information criterion (BIC), and an empirical approach using *a priori* information. A modified ridge regression estimator, which shrinks effects of confounders toward zero, was also considered. For each criterion, the uncertainty in the estimated exposure effect was assessed using bootstrap simulations where confounders were selected in each sample.

These methods were illustrated using data on mercury neurotoxicity in Faroe Islands children. Point estimates and standard errors of mercury effects on confounder-sensitive neurobehavioral outcomes were calculated for each selection procedure.

RESULTS: The full model and the empirical a priori model showed approximately the same precision and these methods were (slightly) inferior only to modified ridge regression. Lower precisions were obtained using backward elimination with a low cut-off level, the BIC and the CIE.

CONCLUSIONS: Standard analysis ignores model selection uncertainty and is likely to yield over-optimistic inferences. Thus, the traditional backward elimination procedure with p=5% should be avoided. If data-dependent procedures are required for confounder identification, we recommend that inferences are based on bootstrap statistics to describe the selection process.

KEY WORDS. Confounding Factors (Epidemiology), Regression Analysis, Statistical Models

Introduction

In observational studies, exposure values are not assigned randomly to the study subjects. Therefore, exposed and unexposed subjects are likely to differ on a number of variables. If some of these variables are affecting the outcome, then the crude relation between exposure and outcome may give a distorted (confounded) reflection of the causal exposure effect. The control of confounding factors has been one of the central issues in epidemiological research, and adjustment is routinely achieved by stratification or by applying some sort of multiple regression analysis.

The important question now is how the investigator decides which of the potential confounders to control for and which to ignore. Often prior knowledge about population relations is weak and the data is used in the confounder identification process. Unfortunately, no standard procedure is fully satisfactory. One approach (backward deletion) is based on stepwise testing of the effects of the potential confounders on the outcome, while another (change-in-estimate) removes potential confounders as long as the exposure effect does not change too much. Despite the frequent use of such automated techniques, very little formal knowledge is available about the impact of the selection process on the subsequent analysis of the exposure effect. Results from simulation studies of the simple situation, where only one potential confounder is present, seem to favor the change-in-estimate method over methods based on significance testing,[1,2] and other simulation studies indicate that forward selection procedures are of limited value in epidemiology.[3] Results from the related problem of "best subset selection" suggest that precision is overestimated, if inference is based on a model selected using stepwise significance testing.[4,5] Although there is a widespread awareness of this fact, the selection process is almost always ignored in the final analysis, and inferences are made as if the selected model was given a priori. Breiman described this routine procedure as a "quiet scandal".[6]

In this paper, we compare different strategies for confounder selection using data from an epidemiological study performed in the Faroe Islands to investigate the adverse

health effects of prenatal mercury exposure. Methylmercury is a common contaminant in seafood and freshwater fish. While adverse effects have been unequivocally demonstrated in poisoning incidents, the implications of lower-level exposures in fish-eating populations have been controversial.[7] The original analysis of the Faroese data showed adverse effects of prenatal mercury exposure on childhood cognitive development,[8] while a study carried out in the Seychelles did not report any significant effects.[9] In 1998 the White House therefore arranged a workshop to assess the quality of the main mercury studies. It was concluded that the Faroese study had chosen an appropriate approach to confounder identification and adjustment.[10] However, further analysis were outlined including adjustment for new potential confounders. Because of the emphasis on residual confounding and the public-health implications, these variables have been included in advanced analyses presented below. The mercury effect is estimated using conventional confounder selection criteria as well as the method originally used by the Faroese study group.[8] Furthermore, adjusted precision estimates, which take the confounder selection process into account, are calculated using the bootstrap method.

Subjects and Methods

The Faroese Mercury Study

A birth cohort of 1022 children was generated in the Faroe Islands during 1986 and 1987 and is being studied prospectively to examine the possible adverse effects of prenatal exposure to methylmercury. The Faroese population is exposed to methylmercury mainly through consumption of contaminated pilot whale meat. Information about the children's prenatal exposure was obtained by measuring mercury concentrations in cord blood. Just before school entry (i.e. in 1993-1994), the children underwent a detailed neuropsychological examination. A total of 917 children were given neuropsychological tests reflecting different domains of brain function. Of the neuropsychological tests administered to all the children, the Boston Naming Test showed the strongest association with prenatal mercury exposure. The short-term recall on the California Verbal Learning Test (CVLT) showed a weaker association,

with a p-value just below 0.05.[8] In the present paper, we focus on these two outcome variables to illustrate how the estimate of the mercury effect depends on the regression model.

An important reason why the exposure-response relation may be confounded in this study is that, in the capital area of Torshavn, the consumption of pilot whale meat was below the Faroese average, but at the same time this area also provided easier access to education and day care. Here we shall consider the following list of possible confounders. *Demography:* The child's sex and age are obvious predictors of development status. It was also taken into account whether or not the child was living with his or hers parents and whether the child was living with younger or older siblings. *Health:* Major medical risk factors for neurobehavioral dysfunction obtained at birth (i.e. low birth weight, small-for-date, and history of head trauma and meningitis) were combined into a single risk parameter. Birth weight, gestational age, and short nursing may also affect childhood development. *Examination:* Some children had to travel by ferry to the examination site, and whether the child was tested in the morning or in the afternoon was also recorded. *Maternal:* A few (41) of the children have a mother born in Denmark, which may affect language skills and thereby affect test scores. Maternal intelligence was measured by her score on Raven's Progressive Matrices. Maternal age at parturition and maternal smoking habits during pregnancy may also predict childhood abilities. *Socioeconomic:* For the socially homogeneous Faroese society, we used vocational or professional education of each parent, and the employment status of the father, as indicators of social background. Furthermore, children in day-care may have an advantage over other children. *Residence:* A dichotomous covariate (*Town7*) was considered which indicated whether or not the child was living in one of the Faroese towns (Torshavn, Klaksvik or Tvaeraa) at the time of examination.

These parameters were selected on the basis of prior knowledge of potential influence on the outcome variables, as considered in the light of the epidemiological setting

in the Faroe Islands. Most of these variables were thought to be weakly related to mercury exposure, which depends on local and variable whale meat availability and personal food preferences, rather than, say, socioeconomic factors. This list of covariates includes the variables previously considered in the original analysis,[8] but has been extended with parameters that reflect possible differences between the major towns with more than 2,000 inhabitants and the smaller fishing villages. Some of these children had to travel longer by ferry to get to the clinic and could have been tired from the travel. We also included the time of day when the testing took place. Table 1 shows the bivariate association between the mercury concentration in the cord blood and each of the potential confounders.

Confounder selection strategies

The effect of mercury exposure after correction for the confounders is determined by multiple regression analysis. The cord blood mercury concentrations showed a skewed distribution and they were logarithmically transformed mainly to avoid that a few highly exposed children became overly influential in the estimation of the exposure effect. With 20 potential confounders and one exposure variable, the full model includes more than 20 nuisance parameters, in addition to the parameter of interest. To gain power in the estimation of the mercury coefficient, standard statistical procedures prescribe identification and removal of any unnecessary covariates.[11] Several confounder selection methods have been suggested, but the inferential properties of these strategies are still poorly known, and an optimal procedure for confounder selection has not been identified. We, therefore, compare different variable selection methods for estimation of the effect of prenatal mercury exposure. Because the aim is to estimate the exposure effect, we have restricted the selection problem to models including the exposure variable.

In the original analysis of the Faroese data, Grandjean et al. developed an *ad hoc* criterion for confounder selection, combining information across different outcome variables.[8] According to this method, the child's sex and age in addition to the ma-

ternal Raven score were considered obligatory confounders for all outcome variables. Additional confounders were selected as follows: for each neuropsychological test score, important predictors were identified using backward elimination (adjusted for the obligatory covariates) with p=0.10. Predictors that were important for more than 3 outcomes (out of 17) were then included in the final regression model for all outcomes. The results of this method, here denoted PGS (Philippe Grandjean selection), are compared with the results of four conventional selection methods.

Backward Elimination (BE): This procedure is based on significance testing, and, despite strong criticism,[12,13] it is still the default solution for model selection. The starting point is the full model adjusting for all possible confounders. Then, one covariate at a time is deleted in a stepwise fashion, at each step deleting the covariate with the highest p-value. The deletion process stops when the p-value of the least significant covariate is below a certain cut-off level. Thus, this procedure rests on the premise that a given covariate is not a confounder if it does not affect the response. It has been argued[1,12] that a significance test of the covariate effect places the burden of proof in the wrong direction, i.e., a covariate is only accepted for control, if its effect on the response is significant. According to this view, the backward elimination process may yield biased effect estimates due to under-selection of important confounders, unless the cut-off is set much higher than the conventional level of 5%. We have therefore investigated this method for cut-off levels of 5%, 10% and 20%.

Change-in-Estimate (CIE): As in backward elimination, this procedure deletes the potential confounders in a stepwise fashion with the full model as the starting point. At each step, the covariate that causes the smallest change in the exposure effect estimate (compared to the full model estimate) upon deletion is removed. The process stops when deletion of each of the remaining variables causes a relative change of more than a given cut-off level, which is usually set at 10%. The idea here is that if the most important confounders are taken into account, then the full model estimate will have a low bias (though possibly a high variance). Whether or not a given

covariate should be considered an important confounder is decided directly from the change in the target parameter, caused by not adjusting for the variable at hand. This procedure has been recommended over the p-value based methods.[1,14] However, few formal results on the statistical properties of the CIE-method are available. In our investigation, we used the recommended cut-off value of 10%.

Akaike's Information Criterion (AIC): According to the AIC, the best model is the one with the minimum value of $-2 \cdot \log(L) + 2 \cdot k$, where L denotes the maximum value of the likelihood function and k is the number of free parameters in the model. Akaike derived this procedure while trying to identify the optimal model for prediction given that the prediction error is determined by the expected Kullback-Leibler distance between the data generating density and the estimated density.[15] Burnham and Anderson strongly recommended the AIC for model selection in biological sciences, mainly because this principle is not dependent upon the unrealistic assumption that the true model is one of the models considered.[16] However, on the subject of prediction-based selection methods, Greenland appropriately stated that "a good rule for a prediction problem may be a poor rule for causal analysis",[13] underlining that the problem of "best subset selection" is not equivalent to the problem of confounder identification.

The Bayesian Information Criterion (BIC): The BIC selection method is similar to the AIC, except that here (minus twice the log of) the likelihood function is penalized using the term: $k \cdot \log(n)$, where n denotes the number of observations. Thus, in studies with $n > 7$, larger models are more heavily penalized by the BIC than by the AIC. The BIC was first developed by Schwarz as an asymptotic solution to Bayesian model selection.[17] Rissanen later motivated the BIC from a coding theoretical point of view,[18,19] while Dawid derived an (asymptotically) equivalent selection criterion based on the predictive powers of the proposed models.[20,21] Contrary to the other selection methods considered here, the BIC is consistent; that is, if a sequence of nested models is proposed, and the true model is one of them, then the BIC will

estimate the dimension of the true model consistently. Thus, using the BIC method the probability of under-fitting or over-fitting will converge to zero as the number of observations increases.

Ridge Regression: Rather than selecting between covariates, ridge regression uses all predictors, but shrinks their effects toward zero.[22] This approach can regarded as a Baysian solution to regression analysis, where regression coefficients a priori are considered to be independent normally distributed with mean zero and variance σ_B^2. The extent of shrinkage is controled by a parameter θ, which is equal to σ^2/σ_B^2, where σ^2 is the residual variance of the response given the covariates. From a Bayesian viewpoint, a natural choice of θ is given by $\theta^* = \widehat{\sigma}^2 \cdot p/\Sigma_i \widehat{\beta}_i^2$, where p is the number of covariates and $\widehat{\sigma}^2$ and $\widehat{\beta}_i, i = 1, ..., p$ are estimates of the full model. In typical applications, study variables are standardized, and the shrinkage parameter is the same for all covariates. Thus, ridge regression does not distinguish exposures from confounders. We therefore considered a modified version, where only the confounder effects were shrunken. This estimator can be viewed as an empirical Bayes estimator with a infinite prior variance for the exposure effect, while the prior variance for the confounders is estimated by $\widehat{\sigma}^2/\theta^*$.

Bootstrap Analysis

When the confounder selection process is based on data, a two-stage estimator is used in the estimation of the exposure effect: first, the confounders are identified and then the exposure regression coefficient is calculated in the selected model. For each of the selection criteria described, the statistical properties of the corresponding composite estimator were explored using bootstrap simulations.[23] In regression analysis the bootstrap can be applied in various ways. When studying model uncertainty, the *non-parametric* bootstrap procedure may seem to be the most natural choice, because this method is not dependent on one of the models being true. In this simple approach, the 917 vectors consisting of covariate and response values of the Faroese children are re-sampled with replacement. In each bootstrap sample, the

confounder selection criteria are applied and the mercury effect is estimated in final models. Statistical properties of the selection criteria can then be determined from the empirical distribution of the effect estimates. All bootstrap investigations were based on 10,000 re-samples of the Faroese data. This number is in agreement with recommendations of Burnham and Anderson,[16] who throughly investigated the non-parametric bootstrap as method for incorporating model uncertainty into statistical inference.

The resampling was not restricted to complete cases. The complex PGS criterion is based on the results of 17 different outcomes. Restricting data to children with complete information on 20 potential confounders *and* 17 response variables would lead to a an unacceptable reduction of the available data. Although all children are re-sampled to obtain comparable results between the selection criteria in each sample, the calculations are restricted to children with complete information on the all potential confounders and the response variable under investigation (i.e. the CVLT or the Boston Naming Test). However, in the confounder identification part of the PGS method, each of the 17 backward elimination processes were based on children with complete information on the all potential confounders and the response variable in question.

Using the non-parametric bootstrap, data are re-sampled from the empirical distribution of the observations. Thus, no model assumptions are exploited in the resampling, which means that this method may be robust to mis-specifications in the regression models, such as heteroscedasticity of error terms and non-linearity in the mean terms.[24] A disadvantage of this approach is that the matrix of covariate values is not constant in different bootstrap data sets. This variation typically results in conservative estimates of variances. However, even in moderately large data sets, this effect is likely to be unimportant.[24]

For one of the Faroese outcome variables, the nonparametric bootstrap yielded a vari-

ance estimate which was a little lower than expected for the full model estimator. In further calculations, the *parametric* bootstrap was therefore used to investigate the robustness of the conclusions based on the non-parametric approach. In the parametric bootstrap, a new outcome value is simulated for each child from the distribution estimated in the full model analysis of the original data set. This is done by first calculating the expected value for each observation based on the full model. Then, a normally distributed residual with a variance identical to the residual variance observed in the original data is simulated, and the new outcome value is given as the sum of the expected value and the residual. Thus, using this approach, the matrix of covariate values is constant, but the estimated variances are dependent on the appropriateness of the full model.

Results

For each of the potential confounders, Table 1 shows the (bivariate) association with the mercury exposure. The strongest associations are seen for *Ferry, Mother Faroese*, and *Town7*, but associations are also significant (at the 5%-level) for *Older sib, Day-care, Maternal Raven*, and *Maternal education*. Most of these associations are the result of low consumption of whale meat in the capital of Torshavn. In a multiple regression analysis with the mercury exposure as the dependent variable and all potential confounders as independent variables, 13.4% of the exposure variation was explained. Thus, although some of the exposure-covariate associations are highly significant, this study has rather limited multicollinearity problems for estimation of the mercury exposure effect. In other words, variation of mercury exposure is poorly explained by variables that may affect child development.

Naive Analysis

In this section, the differences between the results of the selection strategies are described, while ignoring the fact that the selection process may affect the statistical properties of the final model estimates. The selection criteria were first applied to the scores on the CVLT. Table 2 shows the covariates that were included in the final

model, while Table 3 gives the main results of the final model inference. For this outcome, the selection criteria introduce important differences in the subsequent inference on the effect of prenatal exposure to mercury. For the BIC and BE $p = 0.05$, it is estimated that a child loses almost 0.6 points per 10-fold increase in the mercury concentration (p=0.017). This effect is 17% stronger than the full model estimate, which has a p-value just above 5%. The deletion of $Town7$ is the main reason for the de-attenuated mercury coefficients for these criteria. Children living in towns tend to do better so when this variable is excluded this advantage is attributed to having a low mercury exposure, because children in towns had lower mercury concentrations at birth (Table 1).

The CIE method eliminates 17 covariates, which is more than for any of the other criteria. Only when using this criterion the covariates *Exam time*, *Paternal employment*, and the child's age are excluded. These covariates are all strong predictors of the CVLT score ($p<0.0001$ for *Paternal employment* - children of employed fathers do better), but because they are weakly associated with the exposure variable (Table 1), deleting them causes only a slight change in the target parameter estimate. However, if the aim of the selection process is to increase precision in the estimated exposure effect, then strong predictors of the outcome not related to the exposure should not be excluded. This is illustrated by the fact that, for the CIE method, the (naive) standard deviation of the mercury effect is *higher* than the corresponding value in the full model.

For the Boston Naming Test (Table 4 & Table 5), the final model inference is less dependent on the selection criterion. No matter which method is used, a highly significant mercury effect is obtained. The mercury coefficient varies from -1.837 to -1.625 with p-values that are below 0.2%. Again, the BIC yields the strongest effect and again this is due to the fact that this criterion is the only one to exclude $Town7$. Surprisingly, the criterion which resembles the BIC the most, BE p=0.05, yields the weakest mercury effect. However, in addition to controlling for $Town7$, this criterion

eliminates the covariate indicating whether the child has any older brothers or sisters. Both these decisions are associated with an attenuation in the mercury coefficient.

Overall, fewer covariates are excluded for the Boston Naming Test, but the CIE method again eliminates 17 of the potential confounders. It may seem surprising that this criterion is the only one to exclude *Maternal education* and *Day-care*, which are strong predictors of the outcome (selected by the restrictive BE p=0.05) and also clearly associated with the mercury exposure (Table 1). However, *Town*7 is included for control by the CIE. When corrected for this variable, the associations between the exposure and the two potential confounders become less strong, and their deletions are associated with changes in the target parameter below 4%.

Incorporation of model selection uncertainty

From the naive standard deviations of the final model mercury coefficients, it appears that the selection methods overall have succeeded in increasing the precision through the variable deletions. As could be expected, this tendency is strongest for the prediction based methods and conventional backward elimination, which are designed to provide a model with a low sum of squared residuals. However, when the uncertainty in the data-dependent selection process is taken into account, the results are less favorable for the selection strategies. For both outcomes, the bootstrap standard deviation of the full model estimate is the third lowest and only slightly higher than that of the best selection criterion. This means that there is no justification for reducing the full model: the precision of the target parameter is not increased, but bias may be introduced after deletion of real confounders. Thus, in these data, the mercury effect should be assessed in the full model where the effect estimate is highly significant for the Boston Naming Test and on the verge of conventional 5% statistical significance for the CVLT. Note that this conclusion is reached even though the selection criteria agree (especially for the CVLT) that a large part of the potential confounders could have been left out of the analysis.

The other main finding here is that, for most of the criteria, there is a satisfactory agreement between the naive standard deviation and the corresponding value obtained using the bootstrap. Thus, although the final model has been chosen from a set of no less than $2^{20} = 1,048,576$ possible models, the amount of over-optimism in naive precision estimates is not critical.

The performance of the modified ridge regression estimator suggests that instead of deleting potential confounders it is better to keep all of them, and shrink their estimated effects. First of all, this method yields mercury effects which are in close agreement with the least squares estimates of the full model, indicating that ridge regression is nearly unbiased. Secondly, this estimator provides the lowest bootstrap standard errors for both outcomes. Thus, in this analysis ridge regression produced the best exposure effect estimator. However, like the selection procdures this method suffers from the fact that estimation precision is overestimated by the naive standard errors. For ridge regression, this is not a result of selection uncertainty, but the bias occurs because the uncertainty in choice of the shrinkage parameter θ is not taken into account.

The approach used by Grandjean et al.[8] seems to provide a reasonable alternative to full model inference. The PGS method yields mercury effects which are close to the full model estimate, and its bootstrap standard error was beaten only by ridge regression for the Boston Naming Test, while it came in fourth for the CVLT. In addition, a reasonable agreement between the naive and the bootstrap standard errors indicates, that the (naive) PGS inference has not become overly optimistic as a result of the data-driven model selection.

Differences between the precisions of the selection criteria are generally small, but some tendencies are clear. Based on the naive standard deviation, BE p=0.05 and especially the BIC appear to provide the most precise estimation. However, when the selection process is taken into account, the opposite result is obtained, thus indi-

cating that, in addition to providing the most variable estimators, these criteria are also associated with the largest amount of over-optimism in the final model inference. Together with the fact that both methods seem to have induced a substantial amount for bias (for the CVLT), this finding illustrates that the BIC and the BE p=0.05 are not appropriate for confounder identification. In agreement with recommendations by Dales and Ury,[12] it is seen that the statistical properties of the BE method are better if the level of significance is increased to 20%.

By definition, the CIE does not introduce much bias. However, because this method may exclude strong predictors of the response, it may yield an imprecise estimate of the exposure effect. This is the case for the Boston Naming Test, where the CIE is associated with the largest bootstrap standard deviation. However, for the CVLT, the CIE is even better than the full model. This discrepancy between CIE results indicates that the precision of the CIE may be strongly dependent on the specific circumstances in which it is used. Thus, although the CIE may attack the problem of confounder identification in a more direct way than the methods based on significance testing, BE with a p-value of 20% would seem to provide a better option.

For the full model, a close agreement was expected between standard deviations obtained using the naive estimator and the non-parametric bootstrap. Such an agreement was seen for the CVLT. However, for the Boston Naming Test, the bootstrap estimate was somewhat lower than the naive estimate. The parametric bootstrap was therefore applied to investigate the robustness of the findings in the previous section to the choice of resampling distribution. Table 5 gives the estimated standard deviations using the parametric bootstrap. Since the full model is true for the re-sampled data sets, it is no surprise that the full model standard deviation is now close to the naive result. For the selection criteria, similar increases are seen, and the results of the non-parametric bootstrap are therefore confirmed. Thus, ridge regression and the PGS method are again seen to provide the most precise effect estimate, while the BIC and especially the CIE are poorest. Contrary to the non-parametric approach,

the parametric results indicate that a small amount of power may be gained using the AIC or BE with p=20% compared to the full model inference. However, the main conclusion is unchanged: in these data, the possible increase in power obtained through model reductions is too small to justify the use of automatic variable selection procedures.

Discussion

In epidemiology, the researcher is often faced with the seemingly simple task of estimating the effect of one variable (the exposure) on another (the response). However, this task is complicated if inference is drawn based on observational data, because then the effects of an unknown set of confounding variables have to be taken into account. Prior to the statistical analysis, it may be possible to develop a set of potential confounders, which is assumed to include the true confounders. As biological understanding is typically limited, the number of anticipated confounders may be large. This means that the full model with all the potential confounders contains a large number of nuisance parameters. To many investigators, it may seem unappealing to base the exposure inference on a model where some parameters are clearly insignificant. Instead, the model is reduced usually by using one of the subset selection criteria described in this paper, whereupon the exposure effect is estimated in the final model. The results presented here indicate that often it would have been better to assess the exposure effect in the full model. Contrary to what is often indicated by the naive estimates of precision, the model reductions may *increase* estimation variability in addition to introducing biases in exposure regression coefficient. This finding is in agreement with Raab's simulation results on the statistical properties of forward selection procedures.[3]

This case study does not document that full model inference is always superior. In studies with fewer subjects or more potential confounders with stronger associations to the exposure parameter, it may be possible to gain an important amount of power through variable exclusions. However, in such studies model selection uncertainty

will be stronger, and naive standard errors will be more heavily underestimated than was the case in the Faroese mercury data. Thus, it will be even more important to adjust for selection uncertainty in the final model inference. Because of the complex nature of the composite selection estimators, no firm theory is currently available to perform such adjustments. In this regard, it should be noted that Hjort and Claeskens recently presented asymptotic results for the bias and precision of a certain class of composite estimators.[25] However, these results depend on an assumption of "local misspecification", which may not be satisfied, and results on the properties in finite samples are not provided. Therefore, the bootstrap approach constitutes the obvious choice for incorporation of the confounder selection process into the final inference. With today's high-speed computers, this method can be applied quite easily, thereby leading to better inference regarding the effects of the exposure.

In a given study, it may be helpful first to compare the standard error of the effect estimate in the full model to the naive standard error estimate in the selected model. If a substantial variance reduction is not seen, then the full model should be used for inference. However, if the naive standard error in the selected model is clearly lower and if the effect estimate is robust to the variable deletions, this finding would indicate that power can be gained from covariate deletions. If a selection criterion is used, then bootstrap simulations should be conducted to quantify the model uncertainty, and thereby achieve a correct assessment of the significance of the effect of the exposure.

An advantage of the conventional selection criteria is that they have been incorporated in many statistical software packages, which will facilitate the application of the bootstrap. Of these methods, the BIC and backward elimination, with the traditional level of 5%, have been shown to be poorly suited for confounder identification. With these criteria, the risk of deleting important confounders is high, and the estimation uncertainty will be underdetermined. The epidemiological CIE also takes into account the covariate-exposure relation, when a potential confounder is assessed.

In this way, bias in the exposure coefficient is limited, but the final estimate may have a relatively large variance. Backward elimination with a p-value of 20% seems to provide a better estimation. In the original presentation of the Faroese mercury results, a different solution was applied. Prior information was used to identify those confounders that were mandatory, although some of them turned out to cause very minimal confounding in this particular study. Because several outcome variables were available, and because parameters that acted as confounders in regard to one outcome would also be expected to cause confounding with other outcomes, the empirical data were used to generate a "consensus" list of confounders. The analysis presented here indicates that this approach yields a nearly unbiased estimate of the mercury effect. Furthermore, the results of the non-parametric and the parametric bootstrap simulations show that, although naive PGS inference may be associated with some optimism, the mercury effect cannot be explained as an artifact caused by the data-driven model selection process.

As an alternative to variable selection, shrinking of the confounder effects should be considered. Here this was achieved with a modified version of ridge regression, which excluded the exposure parameter from shrinking. The performance of this estimator will depend on the choice of shrinkage parameter. If this parameter is low, the exposure effect estimator will be almost identical to full model analysis, while a high degree of shrinkage will correspond to excluding all confounders. We used a Bayesian estimate for the shrinkage parameter and the corresponding effect estimator appeared superior to all others considered. This method therefore deserves more attention in epidemiology. In a slightly different setting, Greenland advocated for the use of empirical Bayes estimators like this, although he refrained from estimating the shrinkage parameter.[26] An obvious modification of the method would be to group potential confounders according to prior biological importance, and then only to shrink the effects of the less important variables. In this way, the risk underestimating important confounder effects may be limited, and the effect of the exposure of interest may therefore be more accurately assessed.

The results of this paper are based on the assumption that the full model provides an unbiased estimation of the exposure effect. However, this is unlikely to be the case, if the exposure variable or one or more potential confounders are measured with error. It is well known that measurement error in the exposure variable attenuates the dose-response relation. This attenuation depends on the set of selected confounders and will be most severe in the full model, where the variance of the exposure variable given the confounders is minimal.[27] Thus, in this situation, a sub-model with fewer confounders may seem preferable. However, good inference can be drawn only by correcting for the measurement error, and this correction may require the full model.

References

1. Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. *Am J Epidemiol* 1989;129:125-137.

2. Maldano G, Greenland S. Simulation study of confounder-selection strategies. *Am J Epidemiol* 1993;138:923-936.

3. Raab, GM. Selecting confounders from covariates. *J Roy Statist Soc A* 1994;157:271-283.

4. Hjort U. On model selection in the computer age. *J Statist Plann Inference*, 1989;23:101-115.

5. Miller AJ. *Subset Selection in Regression.* 2^{nd} ed. Chapman and Hall; 2002.

6. Breiman L. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction. *J Am Stat Assoc*, 1992;87:738-754.

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

8. Grandjean P, Weihe P, White RF, et al. Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury. *Neurotoxicol Teratol* 1997;19:417-428.

9. Davidson PW, Myers GJ, Cox C, et al. Effects of prenatal and postnatal methylmercury exposure from fish consumption on neurodevelopment: outcomes at 66 months of age in Seychelles child development study. *JAMA* 1998;280:701-707.

10. National Institutes of Health (NIH). *Scientific Issues Relevant to Assessment of Health Effects from Exposure to Methylmercury.*; 1998

11. Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research.* Van Nostrand Reinhold Company; 1982.

12. Dales LG, Ury HK. An improper use of statistical significance testing in studying covariables. *Int J Epidemiol*, 1978;7:373-375.

13. Greenland S. Cautions in the use of preliminary-test estimators. *Stat Med* 1989;8:669-673.

14. Rothman KJ, Greenland S. *Modern Epidemiology.* 2^{nd} ed. Lippincott-Raven; 1998.

15. Akaike H. (1973). Information theory as an extension of the maximum likelihood principle. *Second International Symposium on Information Theory.* Akademiai Kiado, Budapest.

16. Burnham KP, Anderson DR. *Model Selection and Inference.* Springer-Verlag; 1998.

17. Schwarz G. Estimating the dimension of a model. *Ann Statist*, 1978;6:461-464.

18. Rissanen J. A universal prior for integers and estimation by minimum description length. *Ann Statist* 1983;11:416-431.

19. Rissanen J. *Stochastic Complexity in Statistical Inquiry.* Scientific Publication Company; 1989.

20. Dawid AP. Statistical theory, the prequential approach. *J Roy Statist Soc* 1984;147:278-292.

21. Dawid AP. Prequential Analysis, Stochastic Complexity and Bayesian Inference. In: Bernado JM, Berger JO, Dawid AP, Smith AFM, eds. Bayesian Statistics 4. Oxford University Press; 1992;109-125.

22. Draper NR, Smith H. *Applaied Regression Analysis.* 3^{rd} ed. John Wiley & Sons; 1998.

23. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap.* Chapman & Hall; 1993.

24. Davison AC, Hinkley DV. *Boostrap Methods and their Application.* Cambridge University Press; 1997.

25. Hjort LH, Claeskens G. Frequentist model average estimators. *J Am Stat Assoc* 2003;98:879-900.

26. Greenland S. When should epidemiologic regressions use random coefficients? *Biometrics* 2000;56:915-921.

27. Budtz-Jørgensen E, Keiding N, Grandjean P, et al. Consequences of exposure measurement error for confounder identification in environmental epidemiology. *Stat Med* 2003;22:3089-3100.

Tables

Table 1: Percent change in the mercury concentration, associated with a given differnce in each of the covariates.

Covariate	Change in %	95% conf. limit
Age (1 year increase)	6.86	$-10.97; 28.26$
Sex (girl vs boy)	-8.59	$-18.57; 2.63$
Lives w. parents (yes vs no)	17.13	$-3.67; 42.41$
Younger sibs (yes vs no)	-9.93	$-19.87; 1.23$
Older sibs (yes vs no)	16.59	3.23; 31.68
Birth weight (1 kg increase)	5.58	$-5.41; 17.85$
Gestational age (1 week increase)	-0.46	$-4.86; 4.15$
Risk factors for neuro. dysfunc. (yes vs no)	0.31	$-15.28; 18.76$
Short nursing (yes vs no)	8.51	$-7.74; 27.63$
Ferry (yes vs no)	58.71	36.26; 84.86
Exam. time (afternoon vs morning)	5.93	$-5.65; 18.94$
Mother Faroese (yes vs no)	131.66	77.71; 202.00
Maternal raven (10 point increase)	-13.00	$-18.99; -6.56$
Maternal age (1 year increase)	0.82	$-0.26; 1.90$
Maternal smoking (yes vs no)	7.05	$-4.92; 20.53$
Maternal education (yes vs no)	-14.93	$-24.21; -4.51$
Paternal education (yes vs no)	-3.76	$-15.17; 9.18$
Paternal employment (yes vs no)	10.39	$-5.52; 28.98$
Day care (yes vs no)	-17.93	$-26.85; -7.91$
Town7 (yes vs no)	-30.00	$-37.50; -21.60$

Table 2: Excluded covariates in the analysis of the CVLT. For the stepwise criteria (BE, CIE), the numbers indicate the order of the deletions, while excluded covariates are marked with a circle for the prediction based criteria (AIC, BIC) and the PGS method.

	Backward Elimination						
	$p = 0.05$	$p = 0.10$	$p = 0.20$	AIC	BIC	CIE	PGS
Younger sibs	3	3	3	◯	◯	5	◯
Lives w. parents	5	5	5	◯	◯	3	◯
Exam. time						12	
Ferry	11	11			◯	16	◯
Birth weight	9	9	9	◯	◯	8	◯
Short nursing	2	2	2	◯	◯	2	◯
Maternal smoking	4	4	4	◯	◯	9	◯
Maternal age	12	12			◯	14	
Gestational age	1	1	1	◯	◯	1	◯
Mother Faroese	14	14			◯		◯
Older sibs	8	8	8	◯	◯	4	
Town7	15				◯		
Paternal employment						15	
Paternal education					◯	11	
Maternal education	6	6	6	◯	◯	10	
Day care	7	7	7	◯	◯	6	
Risk factors	10	10		◯	◯	7	
Maternal raven							
Sex	13	13			◯	13	
Age						17	

Table 3: Inference on the mercury effect on the CVLT ignoring (naive analysis) and accounting for (bootstrap) confounder selection uncertainty.

	Naive analysis*				Bootstrap†	
Selection Method	$\widehat{\beta}_{Hg}$	$\widehat{s.e.}$	p-value	Change in %‡	mean§	$\widehat{s.e.}$¶
Full model	−0.4983	0.2570	0.0529	-	−0.4948	0.2553
BE $p = 0.20$	−0.5020	0.2555	0.0498	−0.74	−0.5012	0.2597
BE $p = 0.10$	−0.4795	0.2486	0.0542	3.77	−0.5055	0.2631
BE $p = 0.05$	−0.5842	0.2432	0.0165	−17.24	−0.5112	0.2658
AIC	−0.4998	0.2556	0.0509	−0.30	−0.5027	0.2610
BIC	−0.5840	0.2438	0.0168	−17.20	−0.5220	0.2679
CIE	−0.4571	0.2586	0.0775	8.27	−0.4902	0.2551
PGS	−0.4671	0.2493	0.0613	6.26	−0.4740	0.2587
Ridge regression	−0.5004	0.2432	0.0396	−0.42	−0.4961	0.2500

* Results are based on 789 children with complete information

† Number of bootstrap re-samples was 10,000

‡ Relative difference between mercury coefficients in full model and in selected model

§ Empirical mean of bootstraped mercury coefficients

¶ Empirical standard deviation of bootstraped mercury coefficients

Table 4: Excluded covariates in the analysis of the Boston Naming Test. For the stepwise criteria (BE, CIE), the numbers indicate the order of the deletions, while excluded covariates are marked with a circle for the prediction based criteria (AIC, BIC) and the PGS method.

	Backward Elimination						
	$p = 0.05$	$p = 0.10$	$p = 0.20$	AIC	BIC	CIE	PGS
Younger sibs	11				◯	11	◯
Lives w. parents	2	2	2	◯	◯	2	◯
Exam. time	6	6	6	◯	◯	7	
Ferry	3	3	3	◯	◯	9	◯
Birth weight	1	1	1	◯	◯	3	◯
Short nursing					◯	1	◯
Maternal smoking	7	7	7	◯	◯	10	◯
Maternal age	8	8	8	◯	◯	6	
Gestational age	4	4	4	◯	◯	4	◯
Mother Faroese	5	5	5	◯	◯	14	◯
Older sibs	10				◯	5	
Town7					◯		
Paternal employment	9				◯	13	
Paternal education					◯	8	
Maternal education						12	
Day care						16	
Risk factors						15	
Maternal raven							
Sex	12				◯	17	
Age							

Table 5: Inference on the mercury effect on the Boston Naming Test ignoring (naive analysis) and accounting for (bootstrap) confounder selection uncertainty.

	Naive analysis*				Bootstrap†			
					Non-parametric		Parametric	
Selection Method	$\widehat{\beta}_{Hg}$	$\widehat{s.e.}$	p-value	Change in %‡	mean§	$\widehat{s.e.}$¶	mean§	$\widehat{s.e.}$¶
Full model	−1.695	0.5087	0.0009	-	−1.692	0.4949	−1.699	0.5086
BE $p = 0.20$	−1.734	0.4917	0.0004	−2.30	−1.703	0.4950	−1.710	0.5076
BE $p = 0.10$	−1.734	0.4917	0.0004	−2.30	−1.714	0.4972	−1.720	0.5086
BE $p = 0.05$	−1.625	0.4923	0.0010	4.13	−1.725	0.4991	−1.731	0.5107
AIC	−1.734	0.4917	0.0004	−2.30	−1.706	0.4959	−1.712	0.5074
BIC	−1.837	0.4873	0.0002	−8.38	−1.756	0.5011	−1.759	0.5139
CIE	−1.699	0.5069	0.0008	0.24	−1.686	0.5077	−1.677	0.5220
PGS	−1.722	0.4938	0.0005	−1.57	−1.712	0.4932	−1.717	0.5054
Rigde regression	−1.708	0.4846	0.0004	−0.74	−1.703	0.4874	−1.709	0.5033

* Results are based on 782 children with complete information

† Number of bootstrap re-samples was 10,000

‡ Relative difference between mercury coefficients in full model and in selected model

§ Empirical mean of bootstraped mercury coefficients

¶ Empirical standard deviation of bootstraped mercury coefficients