APPENDIX AVAILABLE ON REQUEST

Research Report 152

Evaluating Heterogeneity in Indoor and Outdoor Air Pollution Using Land-Use Regression and Constrained Factor Analysis

Jonathan L. Levy et. al.

Appendix G. Derivation of the Power Expression for Exposure Misclassification Analysis

Note: Appendices Available on the Web appear in a different order than in the original Investigators' Report. HEI has not changed these documents. Appendices were relettered as follows:

Appendix D was originally Appendix A
Appendix E was originally Appendix B
Appendix F was originally Appendix C
Appendix G was originally Appendix D

Correspondence may be addressed to Dr Jonathan I. Levy, 715 Albany St., Talbot 4W, Boston, MA 02118.

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR–83234701 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and therefore may not necessarily reflect the views of the Agency, and no official endorsement by it should be inferred. The contents of this document also have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views or policies of these parties, and no endorsement by them should be inferred.

This document was reviewed by the HEI Health Review Committee
but did not undergo the HEI scientific editing and production process.

Appendix D: Derivation of power expression for exposure misclassification analysis.

Following Vaeth and Skovlund (Vaeth and Skovlund 2004), the asymptotic variance estimator for the health effect of exposure in a simple logistic regression is

$$Var(\hat{\beta}_X) = \frac{1}{\sum_i p_i(1-p_i)\left[\sum_i h_i x_i^2 - \left(\sum_i h_i x_i\right)^2\right]}$$

where X is the exposure, $\hat{\beta}_X$ is the estimated health effect for exposure X, $h_i = p_i(1-p_i) / \sum_j p_j(1-p_j)$, and p_i is the probability of the event for the ith observation. Assuming that the true health effect (β_X) is small, in which case p_i is nearly constant and $h_i \approx 1/n$, then we have $\mathrm{Var}(\hat{\beta}_X) \approx (\sum p_i(1-p_i)s_X^2)^{-1}$, where s_X^2 is the sample variance of the exposure.

Next we consider surrogate exposure values using W to estimate the health effect in the absence of X, where W would be the predictions from an exposure model. For this regression calibration-type setting, the approximate measurement error relationship between X and W is $X = W + \varepsilon$, where the Berkson-style measurement error, ε, is heteroscedastic and correlated. The heteroscedasticity and correlation are often not extreme, so we can approximately decompose the variance as $s_X^2 = s_W^2 + s_\varepsilon^2$, where s_W^2 is the sample variance of W and s_ε^2 is the residual variance from the regression. If we were to use W directly in place of X to estimate β_X using $\hat{\beta}_W$ (the estimated health effect for surrogate W), and use the asymptotic variance estimator with no adjustment for measurement error, a rough estimate of $\mathrm{Var}(\hat{\beta}_W)$ is $(\sum p_i(1-p_i)s_W^2)^{-1}$. Using the standard variance estimator for $\hat{\beta}_W$ when regressing on the surrogate ignores overdispersion induced by use of the surrogate and underestimates the variance, but when β is not too large, this inflation is relatively minor.

The R^2 from the regression model is $R^2 = 1 - s_\varepsilon^2 / s_X^2 = s_W^2 / s_X^2$, so we can approximate the ratio of the variance under the surrogate to that under the true exposure using the simple rule-of-thumb,

$$\frac{\mathrm{Var}(\hat{\beta}_W)}{\mathrm{Var}(\hat{\beta}_X)} \approx \frac{s_X^2}{s_W^2} \approx \frac{1}{R^2}.$$

A reviewer pointed out that this same relationship can be derived as the asymptotic relative efficiency of the two estimators (Lagakos 1988).

This gives us a simple quantification of the inflation in uncertainty from using the surrogate. For example, for an $R^2 = 0.20$, we would expect the variance to be inflated by a factor of five. In making calculations based on real data, we suggest an estimate of R^2 based on cross-validation to avoid bias from overfitting in estimating predictive power.

Based on this, we can derive an approximate relationship between the power using the true exposure, X, and that using the surrogate, W, solely as a function of the R^2 from the exposure model. Under the asymptotic normality of the maximum likelihood estimator, $\hat{\beta} \sim N(\beta, Var(\hat{\beta}))$, the power to detect a significant positive association using a one-sided test based on the true exposure is

$$Power_X \approx \Pr(\hat{\beta}_X > c\sqrt{Var(\hat{\beta}_X)}) = \Phi(-c + \beta / \sqrt{Var(\hat{\beta}_X)})$$

where $\sqrt{\mathrm{Var}(\hat{\beta}_X)} = \mathrm{s.e.}(\hat{\beta}_X)$ is the standard error, Φ is the cumulative normal distribution function, and c is the appropriate critical value under a normal distribution. Next substitute $\sqrt{\mathrm{Var}(\hat{\beta}_X)} \approx \sqrt{R^2}\sqrt{\mathrm{Var}(\hat{\beta}_W)}$ and use simple algebra to express $\sqrt{Var(\hat{\beta}_W)} = \beta(\sqrt{R^2}(c + \Phi^{-1}(Power_X)))^{-1}$. Plugging this expression into the expression for the power to detect a significant positive association using a one-sided test based on the surrogate, we get

$$Power_W \approx \Phi(-c + \beta / \sqrt{Var(\hat{\beta}_W)}) = \Phi(-c + \sqrt{R^2}(c + \Phi^{-1}(Power_X)))$$

In practice, the p-values reported in standard software are based on two-sided tests, but one would only consider significance when the exposure effect is in the expected direction, so in Equation (15) and Figure 13 in the main body of the report, we report the surrogate power with c=1.96. This approximate relationship indicates that the power based on W increases as a function of the power under the true exposure, as one would expect, but in a nonlinear fashion, with low power for small R^2 values.