American Journal of Epidemiology

Vol. 160, No. 4
Printed in U.S.A.
DOI: 10.1093/aje/kwh211

PRACTICE OF EPIDEMIOLOGY

Monte Carlo Sensitivity Analysis and Bayesian Analysis of Smoking as an Unmeasured Confounder in a Study of Silica and Lung Cancer

Kyle Steenland[1] and Sander Greenland[2]

[1] Department of Environmental and Occupational Health, Rollins School of Public Health, Emory University, Atlanta, GA.
[2] Departments of Epidemiology and Statistics, University of California, Los Angeles, Los Angeles, CA.

Received for publication September 5, 2003; accepted for publication February 27, 2004.

Conventional confidence intervals reflect uncertainty due to random error but omit uncertainty due to biases, such as confounding, selection bias, and measurement error. Such uncertainty can be quantified, especially if the investigator has some idea of the amount of such bias. A traditional sensitivity analysis produces one or more point estimates for the exposure effect hypothetically adjusted for bias, but it does not provide a range of effect measures given the likely range of bias. Here the authors used Monte Carlo sensitivity analysis and Bayesian bias analysis to provide such a range, using data from a US silica-lung cancer study in which results were potentially confounded by smoking. After positing a distribution for the smoking habits of workers and referents, a distribution of rate ratios for the effect of smoking on lung cancer, and a model for the bias parameter, the authors derived a distribution for the silica-lung cancer rate ratios hypothetically adjusted for smoking. The original standardized mortality ratio for the silica-lung cancer relation was 1.60 (95% confidence interval: 1.31, 1.93). Monte Carlo sensitivity analysis, adjusting for possible confounding by smoking, led to an adjusted standardized mortality ratio of 1.43 (95% Monte Carlo limits: 1.15, 1.78). Bayesian results were similar (95% posterior limits: 1.13, 1.84). The authors believe that these types of analyses, which make explicit and quantify sources of uncertainty, should be more widely adopted by epidemiologists.

Bayes theorem; confounding factors (epidemiology); epidemiologic methods; lung neoplasms; Monte Carlo method; occupational exposure; occupations; smoking

Abbreviations: RR, rate ratio; SMR, standardized mortality ratio.

INTRODUCTION

General considerations about uncertainty

Conventional confidence intervals reflect only uncertainty due to random error, where the latter is often idealized as the error due to random sampling of subjects from a hypothetical superpopulation. Such sampling error does not reflect biases that arise from confounding, mismeasurement, or nonrandom selection of subjects. Because the amount of bias from these sources is unknown and is not accounted for by the confidence intervals, these intervals will often understate the uncertainty one should have about the true effect; that is, they will be too narrow. They may also be shifted upward or downward because of the unknown bias.

Consider, for example, retrospective studies of occupational cohorts. Information on smoking is usually not available, and cohorts of blue-collar workers are often compared with the general population via standardized mortality ratios (SMRs). The general population is known to smoke less than blue-collar workers, and an outcome of interest for inhaled occupational toxins is often lung cancer, for which smoking is a strong risk factor. This situation is likely to make smoking an upward confounder of the SMR. The extent of this confounding is uncertain, however, and this uncertainty

Correspondence to Dr. Kyle Steenland, Department of Environmental and Occupational Health, Rollins School of Public Health, Emory University, Atlanta, GA 30322 (e-mail: nsteenl@sph.emory.edu).

is not reflected in conventional confidence intervals or *p* values.

Quantitative assessment of likely sources of bias can provide more realistic estimates of the total or actual uncertainty, especially if the investigator has some idea of the likely effects of bias from these sources. Given this information, the investigator can use sensitivity analyses to illustrate the possible extent of such biases (1, 2). Ordinary or traditional sensitivity analysis estimates what the true effect measure (e.g., the rate ratio) would be in light of the observed data and some hypothetical level of bias, and it produces one or more hypothetically adjusted point estimates for the effect measure of interest (2). While conducting ordinary sensitivity analysis is an improvement over ignoring bias, it can become difficult to summarize results as the number of parameters determining the bias increases, and it usually does not provide a full range for likely bias in the results (3). It is usually conducted only when the bias in the point estimate is thought to be pointing in a specific direction (e.g., upward confounding), and even then it is often limited to assessing how much bias in a specific direction would have been necessary to obtain the observed estimate or lower confidence limit if the null condition were true. Ordinary sensitivity analysis can be improved upon through the use of Bayesian methods or Monte Carlo sensitivity analysis.

Likelihood and Bayesian methods can be used to incorporate uncertainty regarding bias into the results of analyses (4–11). Bayesian methods require that the investigator specify prior distributions (priors) for unknown parameters. In our example, we are interested in the effect of an occupational exposure on lung cancer, but our observed data do not include smoking, a serious potential confounder. In a Bayesian analysis, one might specify priors for 1) the exposure-lung cancer rate ratio after adjustment for smoking; 2) the rate ratio for the effect of smoking on lung cancer, taken from the literature; and 3) the estimated proportions of smokers in exposed and nonexposed populations, taken from surveys. As in conventional analyses, one would then construct a model for the probability of the data given these parameters (i.e., the likelihood function). Finally, using Bayes' theorem, the priors for unknown parameters would be combined with the probability of the observed data to produce a posterior distribution for the parameter of interest (the smoking-adjusted rate ratio).

Bayesian methods can be somewhat involved and are not easy to implement with standard software. Newer software (e.g., WinBUGS) makes these calculations possible, but learning how to use this software and understanding what it does can in itself be somewhat daunting. As a result, analogous but simpler Monte Carlo sensitivity analyses have been proposed to account for likely bias (12–15); these analyses can approximate Bayesian results under certain conditions (8–10). These conditions include specification of a joint prior distribution for the unknown parameters in as much detail as is typically required in Bayesian analyses.

Monte Carlo sensitivity analysis is an expanded version of ordinary sensitivity analysis, which repeatedly reestimates the effect measure of interest (e.g., the smoking-adjusted rate ratio relating exposure to lung cancer) based on the observed data and the priors for bias sources. As in ordinary sensitivity analysis (2), one can "correct" or adjust the lung cancer rate ratio, dividing an observed (unadjusted) rate ratio by a bias factor which represents the confounding effect of smoking. This bias factor is a function of the proportions of smokers in the exposed and nonexposed populations as well as the rate ratio for the effect of smoking on lung cancer. One must assign prior distributions to these unknown parameters. To account for random error, the distribution of the observed (smoking-unadjusted) rate ratio relating exposure to lung cancer is estimated using conventional methods; for example, the log of the observed rate ratio may be taken to be normally distributed with a mean and variance estimated by the usual likelihood procedures. One can then repeatedly sample from the priors for the bias parameters and from the estimated distribution of the observed rate ratio, each time constructing a bias factor and then using this factor to adjust the observed rate ratio for smoking (8–10). Alternatively, the data themselves may be modified to include a simulated smoking variable (14). In either approach, a distribution for the parameter of interest is generated on the basis of repeated sampling from priors followed by adjustment, rather than sampling directly from a posterior distribution as in Bayesian analysis.

To illustrate these ideas, we apply Monte Carlo sensitivity analysis and Bayesian bias analysis to a common problem in occupational epidemiology: an unknown degree of confounding of lung cancer rates by cigarette smoking in occupational cohort studies. Although our example is from occupational epidemiology, the techniques we discuss can be applied to any study in which one has some idea of the possible form and size of bias sources (4–15). We start with a brief review of ordinary sensitivity analysis, which forms the core of our Monte Carlo sensitivity analysis.

Ordinary sensitivity analysis

Ordinary sensitivity analyses hypothesize differences between the smoking habits of blue-collar workers and the habits of the general population, as well as the effects of smoking on lung cancer. One can then estimate the amount of confounding in the observed rate ratio. Epidemiologic sensitivity analysis for confounding was apparently introduced by Cornfield et al. (16), with later developments by Bross (17) and Axelson (18).

Suppose that the observed lung cancer rate ratio for an occupational cohort is 1.60, that ever smokers constitute 60 percent of the occupational cohort but only 50 percent of the general population (the reference group), and that the rate ratio for ever smokers versus never smokers is 19. If I_0 is the lung cancer incidence rate for never smokers, then $I_0(19)$ is the rate for ever smokers, and the expected lung cancer incidence rate in the general population is

$$I_{\text{nonexp}} = I_0(0.5) + I_0(19)(0.5) = I_0[0.5 + (19)0.5] = I_0(10.0).$$

If smoking were the only reason the occupational cohort had an elevated lung cancer risk, the expected lung cancer rate among the exposed workers would be

$$I_{\text{exp}} = I_0(0.4) + I_0(19)(0.6) = I_0[0.4 + (19)0.6] = I_0(11.8).$$

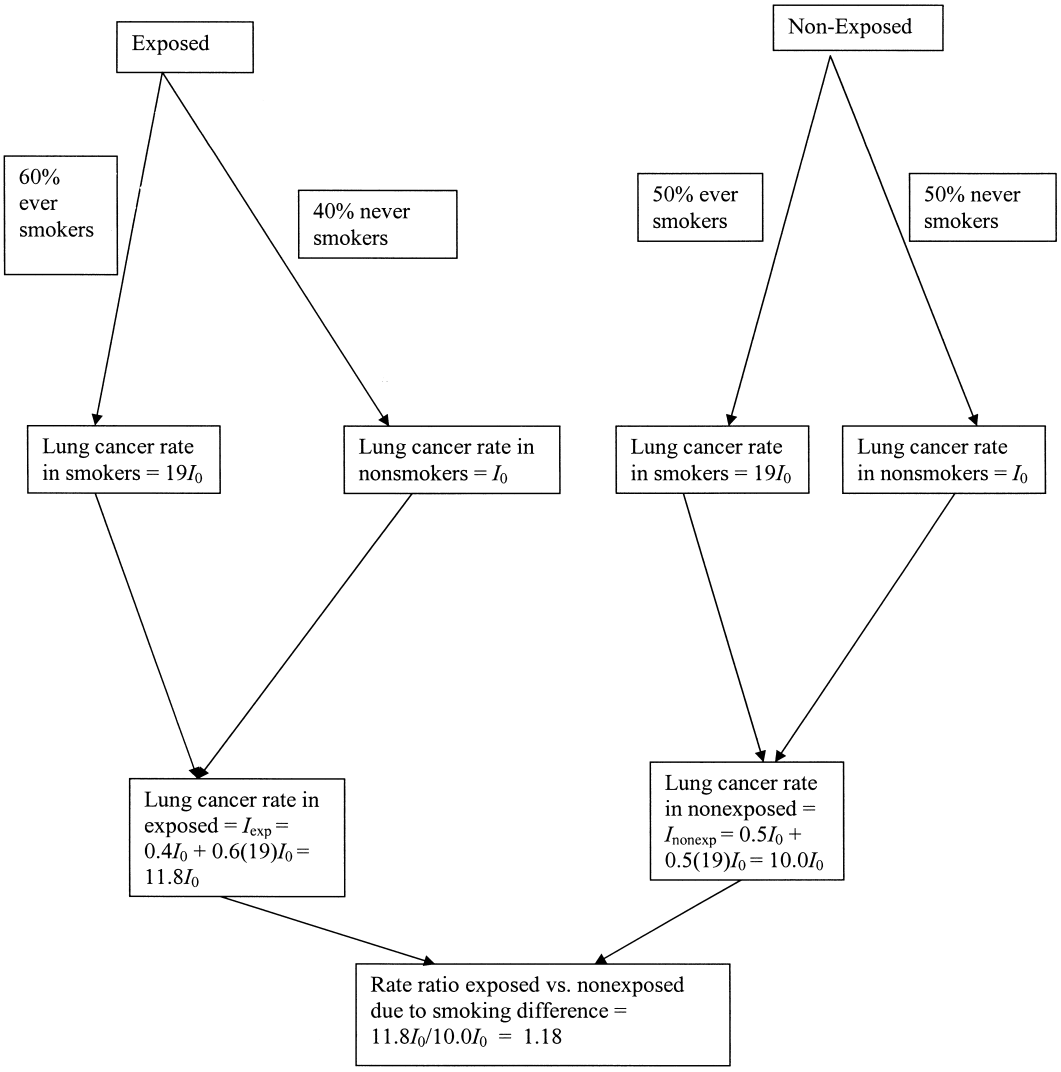

FIGURE 1. Example of confounding of the silica-lung cancer relation by smoking, assuming no effect of silica exposure on lung cancer, in a study of 4,626 US industrial sand workers followed from the 1950s to 1996.

Note that I_{nonexp} and I_{exp} are smoking-weighted averages of the incidence rates for ever smokers and never smokers, but with different weights. The bias factor due to differences in these weights (i.e., the amount of confounding produced by differences in smoking prevalence) is $I_{exp}/I_{nonexp} = 11.8/10.0 = 1.18$. Figure 1 illustrates these calculations.

The bias factor of 1.18 is less than the observed rate ratio of 1.60, suggesting that the observed elevation in the rate among the workers is not due solely to confounding by smoking. The bias factor can also be used to adjust the observed rate ratio: The adjusted estimate would simply be the observed estimate divided by the bias factor, or 1.60/1.18 = 1.36. Many variations on this process have been developed, including formulas for case-control studies and other designs and formulas for correcting for misclassification and selection bias (2).

The above example is typical of actual data on confounding by smoking in occupational studies (18–20), which suggests that differences in smoking between blue-collar workers and the general population probably result in a bias factor on the order of 1.2 for lung cancer.

However, the above corrected point estimate does not account for random error or for uncertainties about the relation of smoking to occupation and lung cancer. To account for the latter, one could consider several scenarios of likely smoking difference (e.g., 70 percent ever smokers in the exposed, 50 percent in the nonexposed) and other smoking-lung cancer rate ratios (e.g., a rate ratio of 15 for ever smoking) and derive a table showing the adjusted rate ratios from each scenario. However, such a table may be cumbersome and difficult to interpret, and may even be misleading because it does not indicate the relative plausibility of different scenarios (3).

One way to expand the scope of ordinary sensitivity analysis is to use Monte Carlo sampling from distributions for confounder prevalences and confounder effects or analogous Bayesian analyses (4–15). Below, we illustrate both of these methods using data from a cohort study of lung cancer among workers exposed to silica (21).

METHODS

The cohort consisted of 4,626 industrial sand workers (99 percent males) exposed to relatively high levels of silica in 18 US plants for an average of 9 years from the 1950s to 1996. A comparison of the cohort with the US population (controlling for age, race, calendar time, and sex) estimated an SMR of 1.60 with 95 percent confidence limits of 1.31 and 1.93 (109 deaths observed vs. 68.1 expected). As is customary, the limits assume that any variability in the observed number of deaths is random and follows a Poisson distribution. Data from a sample of 199 male workers aged 25–64 years from 1987 showed 26 percent never smokers, 40 percent current smokers, and 34 percent former smokers; comparable proportions from 1987 for US males of the same age, based on national survey data (22), were 34 percent, 35 percent, and 31 percent, respectively.

A model for exposure and smoking effects

Let SMR_{unadj} be the SMR for the association of exposure with lung cancer in the study cohort (unadjusted for smoking but adjusted for other covariates, i.e., age, race, and calendar time), and define the following indicator (1 = yes, 0 = no) variables: X_1 = exposure, X_2 = current smoking, and X_3 = former smoking. Assuming that the rate ratios do not vary across covariate levels, the usual log-linear model for the lung-cancer rate ratio (RR) comparing an exposure-smoking-specific group with nonexposed nonsmokers within covariate strata is

$$\text{Exposure-smoking-specific lung cancer RR} = \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3). \quad (1)$$

In this model, $\exp(\beta_1)$ is the rate ratio relating the exposure to the lung cancer rate within specific levels of smoking. It is also SMR_{adj}, the SMR adjusted for smoking and other covariates, the target parameter of interest in this analysis. Furthermore, $\exp(\beta_2)$ and $\exp(\beta_3)$ are the rate ratios for current and former smokers versus never smokers, adjusted for exposure and the other covariates.

Next, let $p_{never,exp}$, $p_{current,exp}$, $p_{former,exp}$, $p_{never,nonexp}$, $p_{current,nonexp}$, and $p_{former,nonexp}$ be the proportions of never, current, and former smokers in the exposed and nonexposed populations within a particular covariate stratum. Under model 1 (and similar to the example in the Introduction), the elevation in the rate of the nonexposed due to smoking (i.e., the rate ratio for ever smokers vs. never smokers among the nonexposed) is a weighted average of the rate ratios for nonexposed never, current, and former smokers, using $p_{never,nonexp}$, $p_{current,nonexp}$, and $p_{former,nonexp}$ as weights:

$$RR_{nonexp0} = p_{never,nonexp} + \exp(\beta_2)p_{current,nonexp} + \exp(\beta_3)p_{former,nonexp}.$$

Similarly, the elevation in the rate of the exposed due to smoking is a weighted average of the same rate ratios for ever smokers, current smokers, and former smokers, but using $p_{never,exp}$, $p_{current,exp}$, and $p_{former,exp}$ as weights:

$$RR_{exp0} = p_{never,exp} + \exp(\beta_2)p_{current,exp} + \exp(\beta_3)p_{former,exp}.$$

As can be seen, any difference between $RR_{nonexp0}$ and RR_{exp0} can only be due to differences in the distributions of smoking among the nonexposed and the exposed. Hence, the elevation in rate due to smoking differences among the exposed relative to the unexposed is $RR_{exp0}/RR_{nonexp0}$, which is the bias due to confounding by smoking:

$$\text{Bias} = RR_{exp0}/RR_{nonexp0} = \frac{p_{never,exp} + \exp(\beta_2)p_{current,exp} + \exp(\beta_3)p_{former,exp}}{p_{never,nonexp} + \exp(\beta_2)p_{current,nonexp} + \exp(\beta_3)p_{former,nonexp}}. \quad (2)$$

This bias term would ordinarily vary across covariate strata with different smoking prevalences. If it were approximately constant across the strata, the smoking-adjusted SMR could be estimated by dividing the unadjusted SMR by that constant bias factor: $SMR_{adj} = SMR_{unadj}/\text{bias}$. For the present example, we will assume that the use of the crude smoking prevalences in equation 2 provides a reasonably accurate estimate of the smoking confounding in SMR_{unadj}. At the cost of added complexity, such assumptions could be avoided by sampling covariate-specific confounder (smoking) prevalences and effects and then using those for covariate-specific adjustment of the exposure-disease relative risk estimate, followed by summarization across covariate levels (9, 10).

Monte Carlo sampling

We used Monte Carlo sampling (comprising 5,000 randomly sampled confounding scenarios) to repeatedly estimate the bias factor (equation 2). Each scenario was based on a prior distribution for the proportions of current and former smokers among the exposed and the US population, independently sampling from a prior distribution for the rate ratio relating smoking to lung cancer. These distributions were constructed from available data and so are largely empirical, the chief subjective judgment being that those data apply to our study. For each scenario, we also resampled the observed rate ratio from its estimated distribution. We then used the estimated bias factor to correct the resampled exposure-lung cancer rate ratio, thus deriving a distribution of smoking-adjusted rate ratios for the exposure-lung cancer relation.

Our prior distribution for the smoking-lung cancer rate ratios was computed from a large cohort study, the American Cancer Society's Cancer Prevention Study II, in which follow-up began in 1982 (23). The rate ratio for current smokers versus never smokers was 23.6 (95 percent confidence interval: 19.6, 28.3), and the rate ratio for former smokers versus never smokers was 8.7 (95 percent confidence interval: 7.2, 10.4). These rate ratios are fairly representative of those observed in other large cohort studies conducted during the same time period. Given these data, we expressed our remaining uncertainty about the smoking log rate ratios β_2 and β_3 in model 1 (equation 1) by giving them a normal distribution with mean values equal to ln(23.6) and ln(8.7) and standard deviations equal to the standard errors of these estimates (estimated from ln(upper confidence limit/lower confidence limit)/3.92, which is 0.094 for both). Using S-Plus (24), we generated 5,000 log rate ratios β_2 and β_3 for current and former smoking from these normal distributions. Because we did not have the original data, which would have allowed us to estimate their correlation, we left these log rate ratios uncorrelated.

Our prior distribution for the smoking prevalences $p_{\text{never,exp}}$, $p_{\text{current,exp}}$, and $p_{\text{former,exp}}$ in the exposed cohort was based on the proportions of never, current, and former smokers (26 percent, 40 percent, and 34 percent, respectively) in the sample of 199 workers. Using standard formulas (25) for normal approximations to the logit of a proportion p, $\ln[p/(1-p)]$, we generated the proportions as follows. First, we sampled 5,000 logits of the proportions of never and current smokers among the exposed workers, $\text{logit}(p_{\text{exp,never}})$ and $\text{logit}(p_{\text{exp,current}})$, from a bivariate normal distribution with respective means logit(0.26) = –1.05 and logit(0.40) = –0.41, standard deviations of $[199(0.26)0.74]^{-1/2} = 0.16$ and $[199(0.40)0.60]^{-1/2} = 0.14$, and a correlation of $-[0.26(0.40)/0.74(0.60)]^{1/2} = -0.48$ (the proportions and hence the logits are negatively correlated; for example, when the proportion of current smokers increases, the proportion of never smokers decreases). Next, we converted the sampled logits back to the proportions $p_{\text{never,exp}}$ and $p_{\text{current,exp}}$ using the conversion formula: proportion = $1/(1 + e^{-\text{logit}})$. Finally, we computed $p_{\text{former,exp}} = 1 - p_{\text{never,exp}} - p_{\text{current,exp}}$. This method of generating the prevalences permits negative values for $p_{\text{former,exp}}$ (which is negative if $p_{\text{never,exp}} + p_{\text{current,exp}} > 1$). However, the negative correlation of the logits of $p_{\text{never,exp}}$ and $p_{\text{current,exp}}$ makes it very improbable that $p_{\text{former,exp}}$ will be negative; if a negative value does occur, that scenario should be discarded. (One can make sure this does not happen by generating the prevalences from a Dirichlet distribution, which is a multivariate generalization of the beta distribution (25), but this distribution may not be familiar to many readers and is not as widely available as the normal distribution; our bivariate normal distribution is derived as an approximation to the logit of a Dirichlet distribution.)

Similarly, our smoking distribution $p_{\text{never,nonexp}}$, $p_{\text{current,nonexp}}$, and $p_{\text{former,nonexp}}$ in the unexposed cohort is based on the proportions of never, current, and former smokers (34 percent, 35 percent, 31 percent, respectively) in the male US population aged 25–64 years in 1987 national survey data (22). These estimates are based on 56,000 such men, but to allow for extra variance due to the clustered design and extra uncertainty due to possible nonresponse and other problems, we used an effective sample size of one fourth of this value, 14,000 (which doubles the standard deviations of the logits). First, we sampled 5,000 logits of the proportions of never and current smokers among the exposed workers, $\text{logit}(p_{\text{nonexp,never}})$ and $\text{logit}(p_{\text{nonexp,current}})$, from a bivariate normal distribution with respective means logit(0.34) = –0.66 and logit(0.35) = –0.62, standard deviations of $[14{,}000(0.34)0.66]^{-1/2} = 0.018$ and $[14{,}000(0.35)0.65]^{-1/2} = 0.018$, and a correlation of $-[0.34(0.35)/0.66(0.65)]^{1/2} = -0.53$. Next, we converted the sampled logits back to the proportions $p_{\text{never,nonexp}}$ and $p_{\text{current,nonexp}}$. Finally, we computed $p_{\text{former,nonexp}} = 1 - p_{\text{never,nonexp}} - p_{\text{current,nonexp}}$.

We then sampled 5,000 unadjusted exposure-lung cancer rate ratios from a normal distribution with a mean equal to the log of the original unadjusted estimate, ln(1.60) = 0.47, and a standard deviation equal to the standard error of that estimate, 0.099. This step adds random-sampling error into the Monte Carlo analysis.

We used the 5,000 generated sets of proportions of never, current, and former smokers in the cohort and in the general population and the 5,000 generated lung cancer rate ratios for current and former smokers versus never smokers to compute 5,000 bias factors using equation 2 above. We then used each of these bias factors to adjust a sampled unadjusted rate ratio. The S-Plus code for our calculations is provided in Appendix 1.

Bayesian analysis

For comparison with the Monte Carlo approach, we conducted a Bayesian analysis (5), in which the observed data are entered into a data model and then combined with prior distributions for the parameters in the model to derive a posterior distribution for the parameters. We performed this analysis using WinBUGS (26), which generates samples of the parameters from the posterior distribution. The WinBUGS code is provided in Appendix 2. The data model specified that the observed number of lung cancer deaths ($n = 109$) was from a Poisson distribution with mean equal to

the expected number of deaths based on the US population (68.1) times the product of the unknown smoking-adjusted rate ratio (RR_{adj}) and the unknown bias factor. To parallel the Monte Carlo sensitivity analysis (which does not use a prior for RR_{adj}), we specified an essentially noninformative distribution (i.e., one with very large variance) for the log of RR_{adj} (a normal distribution with mean 0 and variance 10,000). The bias factor was again calculated as in equation 2. The same informative prior distributions for these ratios and proportions were specified as in the Monte Carlo analysis. After a burn-in of 5,000 observations, we generated 100,000 smoking-adjusted rate ratios from the posterior distribution using WinBUGS.

RESULTS

The median and geometric mean values of the bias factors generated from the Monte Carlo analysis were both 1.12. The 5,000 unadjusted exposure-lung cancer rate ratios were divided by the 5,000 bias factors to generate 5,000 exposure-lung cancer rate ratios adjusted for confounding by smoking. The median and geometric mean values of the adjusted rate ratios were both 1.43, and the 2.5th and 97.5th percentiles (95 percent Monte Carlo limits) were 1.15 and 1.78. Bayesian posterior sampling yielded median and geometric mean values for the 100,000 adjusted rate ratios of 1.44 and 1.43, respectively, with 2.5th and 97.5th percentiles (estimated 95 percent posterior limits) of 1.13 and 1.84—very similar to the Monte Carlo sensitivity analysis.

Both sets of results should be compared with the original point estimate of 1.60 and 95 percent confidence limits of 1.31 and 1.93. Figure 2 shows the estimated distribution of the original (smoking-unadjusted) rate ratios and the distribution of smoking-adjusted rate ratios from the Monte Carlo sensitivity analysis. The Monte Carlo distribution has a mean that has shifted downward by approximately 10 percent relative to the original point estimate and is a bit more spread out: The ratio of the upper limit to the lower limit (a measure of uncertainty) has increased from the original 1.47 (1.93/1.31 = 1.47) to 1.54 (1.78/1.15 = 1.54) for the Monte Carlo limits (and to 1.63 (1.84/1.13 = 1.63) for the Bayesian limits). The conventional limits give the impression that the silica exposure is probably associated with an increase in the lung cancer rate of at least 30 percent and quite possibly more than 90 percent (relative to the US population), whereas the Monte Carlo and Bayesian results give the impression that this increase could easily be less than 20 percent and is unlikely to be more than 90 percent.

DISCUSSION

The results from the Monte Carlo sensitivity analysis are improvements over traditional sensitivity analyses in that they provide a distribution for the smoking-adjusted rate ratio, rather than just a few hypothetical guesses. Of course, this distribution depends on the input distributions, but these are displayed in a precise manner to allow scrutiny, and readers who dispute the specification can rerun the analysis using their own inputs, if they wish.

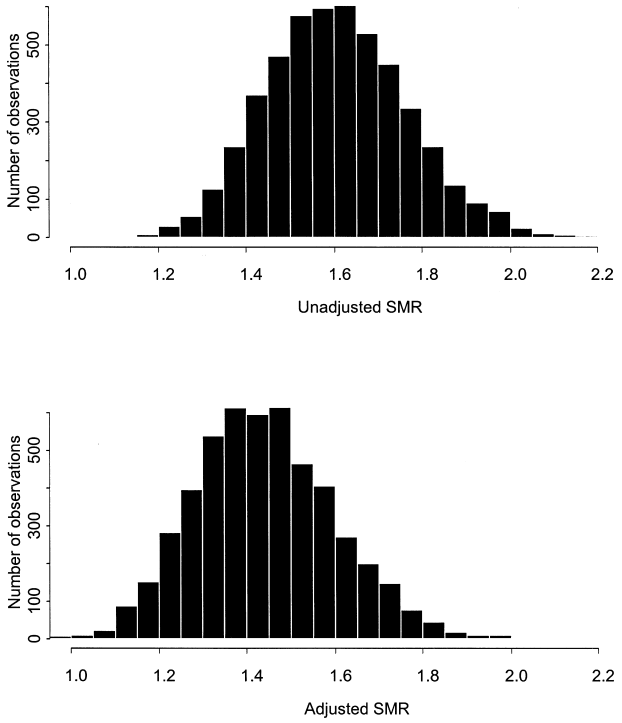

FIGURE 2. Unadjusted and smoking-adjusted lung cancer standardized mortality ratios (SMRs) in a study of 4,626 US industrial sand workers followed from the 1950s to 1996.

Although the Monte Carlo limits were shifted downward, the ratios of upper limits to lower limits (1.54 in the Monte Carlo analysis and 1.63 in the Bayesian analysis) only modestly exceeded the ratio of the 95 percent confidence limits (1.47). This reflects the fact that the correction did not vary widely across scenarios, which in turn reflects the fact that both smoking-lung cancer associations (i.e., current and former smokers vs. never smokers) were sampled with relatively high precision and that the observed differences in smoking habits between the cohort and the US population (based on a sample of workers) tended to be small, which limited the generated bias factors to a small range. In situations where the prior information is less precise, the bias distribution will lead to substantially wider Monte Carlo and Bayesian intervals relative to the conventional confidence intervals (which reflect only random error). Monte Carlo sensitivity analysis and Bayesian bias analysis will also produce much wider intervals when the conventional confidence intervals are narrow, as in large studies, pooling projects, and meta-analyses (9, 10). Another key aspect determining the final results will be the shift (or location) in the final distribution, relative to the original confidence limits. In the example, neither the widening of the interval nor the location shift due to confounding by smoking was large, but together they added up to considerably reduce the lower limit (from 1.31 to 1.15 or 1.13).

No analysis will capture every conceivable source of uncertainty; the goal of bias analysis is to adequately reflect the major sources while avoiding unimportant details. There are some aspects of confounding by smoking that we did not consider. First, we estimated the proportions of never, current, and former smokers in our data at one point in time (1987) based on a sample of the cohort. It is possible to estimate smoking habits in the cohort over time, by using surveys of US blue-collar males over time, and to do the same for the US population. We felt, however, that use of actual data from 1987 for a sample of the cohort was preferable to use of data from surveys of blue-collar workers. Another refinement would use data on the amount smoked rather than categories of never, current, and former smokers. However, we did not have data from our sample of workers with which to estimate pack-years of smoking. Finally, our rate ratios for lung cancer among current and former smokers versus never smokers were chosen from one large cohort of US males. One might instead choose to conduct a meta-analysis of studies of male US smokers; but, as we noted above, the American Cancer Society smoking-lung cancer rate ratios tend to be similar to those observed in other large cohorts studied during the same period.

There may always be other confounders besides smoking. However, because of the strength of its effect and its high prevalence, we think smoking is the main potential confounder for lung diseases, since other potential confounders are either unlikely to have differed in prevalence substantially between our cohort and the US population (e.g., asbestos exposure) or are only modestly related to lung cancer (e.g., diet). A possible healthy worker effect might also have led to some underestimation of the exposure effect.

Uncontrolled confounding by smoking may be the largest bias in our example. Selection bias would be limited because follow-up was relatively complete (95 percent). Exposure (dichotomous) is unlikely to be seriously misclassified: For the worker cohort, exposure status was based on employment with the companies in the study, where silica exposure had been documented; in the general population, only a negligible proportion would be exposed occupationally to silica. If one attempted a dose-response analysis, however, issues relating to measurement error would have to be addressed.

CONCLUSION

It takes an appreciable amount of work to conduct a bias analysis, but such an analysis has the advantage of explicitly and quantifiably taking into account likely sources of bias. Those sources are usually discussed in a more qualitative and abbreviated manner in the Discussion section of most papers, as we did for misclassification and selection bias. If it is very clear that the biases must be small or if the conventional confidence intervals are so wide that nothing could be inferred even in the absence of bias, these discussions may be sufficient. However, if there are concerns that biases may be large, or if the conventional results appear very precise (and thus potentially very misleading if taken at face value), bias analyses can play a crucial role in forming inferences. They help by forcing the investigators to make their judgments about bias sources explicit and precise and by showing the impact these judgments should have on inferences. We thus recommend that bias analysis become a required component of training in epidemiologic methods and that it be implemented whenever conventional confidence intervals that will be used for policy or planning purposes appear decisive or narrow.

ACKNOWLEDGMENTS

Drs. Haitao Chu and Lance Waller kindly provided advice and help with computer code, and Dr. Katherine Hoggatt kindly provided comments on the manuscript.

REFERENCES

1. Maldonado GM, Delzell E, Tyl R, et al. Occupational exposure to glycol ethers and human congenital malformations. Int Arch Occup Environ Health 2003;76:405–23.
2. Greenland S. Basic methods for sensitivity analysis and external adjustment. In: Rothman KJ, Greenland S. Modern epidemiology. 2nd ed. Philadelphia, PA: Lippincott-Raven Publishers, 1998:343–57.
3. Greenland S. The sensitivity of a sensitivity analysis. In: 1997 proceedings of the biometrics section. Alexandria, VA: American Statistical Association, 1998:19–21.
4. Eddy D, Hasselblad V, Schacter R. Meta-analysis by the confidence profile method. New York, NY: Academic Press, Inc, 1992.
5. Gelman A, Carlin JB, Stern HS, et al. Bayesian data analysis. 2nd ed. New York, NY: Chapman and Hall, Inc, 2003.
6. Graham P. Bayesian inference for a generalized population attributable fraction. Stat Med 2000;19:937–56.
7. Little RJ, Rubin DB. Statistical analysis with missing data. 2nd ed. New York, NY: John Wiley and Sons, Inc, 2002.
8. Greenland S. Sensitivity analysis, Monte Carlo risk analysis, and Bayesian uncertainty assessment. Risk Anal 2001;21:579–83.
9. Greenland S. The impact of prior distributions for uncontrolled confounding and response bias: a case study of the relation of wire codes and magnetic fields to childhood leukemia. J Am Stat Assoc 2003;98:47–54.
10. Greenland S. Multiple-bias modeling for analysis of observational data (with discussion). J R Stat Soc (in press).
11. Gustafson P. Measurement error and misclassification in statistics and epidemiology and statistics. New York, NY: Chapman and Hall, Inc, 2003.
12. Lash TL, Silliman R. A sensitivity analysis to separate bias due to confounding from bias due to predicting misclassification by a variable that does both. Epidemiology 2000;11:544–9.
13. Powell M, Ebel E, Schlossel W. Considering uncertainty in comparing the burden of illness due to foodborne microbial pathogens. Int J Food Microbiol 2001;69:209–15.
14. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. Epidemiology 2003;14:451–8.
15. Phillips C. Quantifying and reporting uncertainty from systematic errors. Epidemiology 2003;24:459–66.
16. Cornfield J, Haenszel W, Hammond E, et al. Smoking and lung cancer: recent evidence and a discussion of some questions. J Natl Cancer Inst 1959;22:173–203.
17. Bross ID. Pertinency of an extraneous variable. J Chronic Dis

1967;20:487–95.
18. Axelson O. Aspects on confounding in occupational health epidemiology. Scand J Work Environ Health 1978;4:85–9.
19. Siemiatycki J, Wacholder S, Dewar R, et al. Degrees of confounding bias related to smoking, ethnic group, and socioeconomic status in estimates of the association between occupation and cancer. J Occup Med 1988;30:617–25.
20. Axelson O, Steenland K. Indirect methods of assessing tobacco use in occupational studies. Am J Ind Med 1988;13:105–18.
21. Steenland K, Sanderson W. Lung cancer among industrial sand workers exposed to crystalline silica. Am J Epidemiol 2001; 153:695–703.
22. National Center for Health Statistics. Smoking and other tobacco use: United States, 1987. (Vital and health statistics, series 10, no. 169). Hyattsville, MD: National Center for Health Statistics, 1989.
23. Thun M, Apicella L, Henley J. Smoking versus other risk factors as the cause of smoking-attributable deaths: confounding in the courtroom. JAMA 2000;284:706–12.
24. Insightful Corporation. S-Plus 6.1 for Windows. Seattle, WA: Insightful Corporation, 2002.
25. Bishop YM, Fienberg SE, Holland PW. Discrete multivariate analysis: theory and practice. Cambridge, MA: MIT Press, 1975.
26. MRC Biostatistics Unit. WinBUGS 1.4. London, United Kingdom: Imperial College of Science, Technology, and Medicine, 2003.

APPENDIX 1

S-Plus Code for Monte Carlo Sensitivity Analysis

```
muexp<–c(–1.05, –0.41)
## means for multivariate normal distribution of logit of proportions of never and current smokers among exposed
munexp<–c(–0.66, –0.62)
## means for multivariate normal distribution of logit of proportions of never and current smokers among nonexposed
covexp<–matrix(c(0.026, –0.011, –0.011, 0.021), ncol=2)
## variance-covariance matrix for logits of proportions of never and current smokers among exposed
covnexp<–matrix(c(0.00032, –0.00017, –0.00017, 0.00031), ncol=2)
## variance-covariance matrix for logits of proportions of current smokers, exposed and nonexposed
xexp<–rmvnorm(5000, mean=muexp, cov=covexp, d=2)
## draw from a multivariate normal distribution for logits of proportions of never and current smokers among exposed
xnexp<–rmvnorm(5000, mean=munexp, cov=covnexp, d=2)
## draw from a multivariate normal distribution for logits of proportions of never and current smokers among nonexposed
bcur<–rnorm(5000, 3.16, sd=0.0939)
## prior normal distribution for log rate ratio (RR) for male current smokers versus never smokers, from Cancer Prevention Study II data; RR = 23.6 (95 percent confidence interval: 19.6, 28.3)
bform<–rnorm(5000, 2.16, sd=0.0938)
## prior normal distribution for log RR for male former smokers versus never smokers, from Cancer Prevention Study II data; RR = 8.7 (95 percent confidence interval: 7.2, 10.4)
bexposure<–rnorm(5000, 0.47, 0.1)
## observed exposure-disease log RR and standard error, with no adjustment for smoking; reflects random error
pnev1<–exp(xexp[,1])/(1+exp(xexp[,1]))
## proportion of never smokers in exposed, derived from 1987 sample of exposed cohort, males aged 25–64 years
pnev0<–exp(xnexp[,1])/(1+exp(xnexp[,1]))
## proportion of never smokers in nonexposed, derived from 1989 US survey of males aged 25–64 years
pcur1<–exp(xexp[,2])/(1+exp(xexp[,2]))
## proportion of current smokers in exposed, derived from 1987 sample of exposed cohort, males aged 25–64 years
pcur0<–exp(xnexp[,2])/(1+exp(xnexp[,2]))
## proportion of current smokers in nonexposed, derived from 1989 US survey of males aged 25–64 years
pform1<–1–pnev1–pcur1
## estimated proportion of former smokers in exposed
pform0<–1–pnev0–pcur0
## estimated proportion of former smokers in nonexposed
bias<–(pnev1+exp(bcur)*(pcur1)+exp(bform)*(pform1))/(pnev0+exp(bcur)*(pcur0)+exp(bform)*(pform0))
## bias parameter
bnew<–bexp–log(bias)
## value of RR for exposed versus nonexposed corrected for bias due to confounding by smoking
exp(mean(bnew))
exp(quantile(bnew,probs=0.025))
exp(quantile(bnew,probs=0.50))
exp(quantile(bnew,probs=0.975))
## percentiles of smoking-adjusted rate ratio
mean(bias)
## mean bias parameter
quantile(bias,probs=0.50)
## median bias parameter
```

APPENDIX 2

WinBUGS Code for Bayesian Analysis

```
model
{
pnev1<–exp(xexp[1])/(1+exp(xexp[1]))
# proportion of never smokers among exposed, a function of the logit of this proportion
pnev0<–exp(xnexp[1])/(1+exp(xnexp[1]))
# proportion of never smokers among nonexposed, a function of the logit of this proportion
pcur1<–exp(xexp[2])/(1+exp(xexp[2]))
# proportion of current smokers among exposed, a function of the logit of this proportion
pcur0<–exp(xnexp[2])/(1+exp(xnexp[2]))
# proportion of current smokers among nonexposed, a function of the logit of this proportion
pform1<–1–pcur1–pnev1
# proportion of former smokers among exposed
```

```
pform0<–1–pcur0–pnev0
# proportion of former smokers among nonexposed
bias<–(pnev1+exp(betacur)*(pcur1)+exp(betaform)*(pform1))/
  (pnev0+exp(betacur)*(pcur0)+exp(betaform)*(pform0))
# bias is the bias parameter; it is a function of the proportions
  of never, current, and former smokers among exposed and
  nonexposed and the rate ratios for current and former
  smoking
betaobs<–betanew+log(bias)
# log standardized mortality ratio (SMR) observed
  (confounded) = log SMR adjusted for confounding +
  log(bias)
lambda<–68*exp(betaobs)
# Poisson mean of observed deaths is the expected number of
  deaths times the SMR observed
obs~dpois(lambda)
# observed number of deaths are distributed as a Poisson
  variable with mean lambda
betanew~dnorm(0,0.0001)
# assumed distribution of log SMR true (unconfounded), a
  diffuse prior distribution
# note, WinBUGS uses inverse of variance
betacur~dnorm(3.16,113.65)
# an informative prior for lung cancer rate ratio for current
  smoking, from American Cancer Society data; rate ratio =
  23.6 (95 percent confidence interval: 19.6, 28.3)
betaform~dnorm(2.16,113.65)
# an informative prior for lung cancer rate ratio for former
  smokers, from American Cancer Society data; rate ratio =
  8.7 (95 percent confidence interval: 7.2, 10.4)
xexp[1:2]~dmnorm(muexp[1:2], covexp[1:2, 1:2])
# logits of proportions of never and current smokers in
  exposed, assumed correlated –0.48, multivariate normal
xnexp[1:2]~dmnorm(munexp[1:2], covnexp[1:2, 1:2])
# logits of proportions of never and current smokers in
  nonexposed US population, assumed correlated –0.53,
  multivariate normal
}
list(obs=109, muexp=c(–1.05,–0.41), munexp=c(–0.66,–0.62),
covexp=structure(.Data=c(3.9333208, 2.0603109, 2.0603109,
  48.698258),.Dim=c(2,2)),
covnexp=structure(.Data=c(4409.6728, 2418.2077, 2418.2077,
  4551.9203),.Dim=c(2,2)))
# observed data, a single observation, the observed number
  of lung cancer deaths in the exposed
# assumed means of the logit of proportions of never and
  current smokers among exposed, taken from observed data
  for a sample of the exposed cohort, in which proportions of
  never and current smokers aged 25–65 years were 0.26 and
  0.40, respectively
# assumed means of the logit of proportions of never and
  current smokers among the nonexposed, taken from US
  population data, in which proportions of male never and
  current smokers were 0.34 and 0.35, respectively
# assumed inverse of covariance matrix for logits of propor-
  tions of never and current smokers in exposed, assuming
  these are correlated at –0.48
# assumed inverse covariance matrix for logits of propor-
  tions of never and current smokers in nonexposed,
  assuming these logits are correlated at –0.53
# end
```