Human and Ecological Risk Assessment: Vol. 2, No. 4, pp. 920-938, 1996

Monte Carlo Modeling of Epidemiological Studies

Alexander Shlyakhter, Leonid Mirny, Alexander Vlasov,* and Richard Wilson**
Department of Physics and Harvard Center for Risk Analysis, Harvard University, Cambridge, MA

ABSTRACT

As epidemiologists search for smaller and smaller effects, the statistical uncertainty in their studies can be dwarfed by biases and systematic uncertainty. We here suggest that Monte Carlo techniques are very useful to estimate some of these biases and uncertainties, and perhaps to avoid them entirely. We illustrate this by two simple Monte Carlo simulations. First, we show how often false positive findings, and sometimes false negative findings, can result from 33 differential misclassification of the exposure status. Secondly, we show how a bias, that we call "the binning bias," can be caused if the investigator chooses bin boundaries after he has seen the data. We show how an allowance might be made for such a bias by increasing the uncertainty bounds. This would put the presentation of the results on a par with the presentation in physical sciences where a quantitative estimate of systematic errors is routinely included with the final result. Finally, we suggest how similar Monte Carlo simulations carried out before and during the study can be used to avoid the biases entirely.

Key Words: uncertainty, modeling, epidemiology, bias

INTRODUCTION

The first publication of one of the authors (RW) was an electronic device to produce random numbers for use in statistical experiments on extrasensory perception (Wilson, 1949). The earliest exposure to Monte Carlo analyses was during a visit to Cornell University in 1950 where Robert Wilson was using Monte Carlo methods to follow the development of electromagnetic showers of electrons and photons in a block of lead (Wilson, 1950, 1951). This is a complex problem that is only analytically solvable using simplistic approximations (Rossi and Greisen, 1947). In this case the random numbers were selected by a roulette wheel. Now, 45

* Visitor from Institute of Radiation Hygiene, 197101, Mira Street 8, St. Petersburg, Russia.
** Person for contact: Department of Physics, Harvard University, Cambridge, MA 02138; Tel: (617) 495-3387; Fax: (617) 495-0416; E-mail: wilson@physics.harvard.edu

years later, this problem is routinely solved by the EGS Monte Carlo program, and although the random numbers are routinely selected electronically by computer, the principle is the same.

In the field of risk analysis, it is often necessary to fold a number of independent distributions of quantities that are combined together. When the quantities are multiplied, and the distributions can be approximated by lognormal distributions, a simple analytic solution is possible (Crouch and Wilson, 1981). The result is itself a lognormal distribution with the standard deviation obtained from the standard deviations of the individual distributions by adding in quadrature. When many distributions that are not lognormal, but nonetheless smooth, are folded together, the result approaches closer and closer to the lognormal, making a lognormal approximation adequate for most purposes. However, when different distributions are *added*, this analytic solution is not possible, and a Monte Carlo approach has been used for many years (Crouch, Wilson, and Zeise, 1983). In 1996, Monte Carlo programs are often used for the simpler problem of combining distributions of quantities that are also multiplied together.

A much more powerful and important role of Monte Carlo techniques in science lies in the simulation of complex experiments and observational protocols where all the parameter space is unavailable for measurement. There are often experimental "constraints" on the measurement. In such situations analytic solutions are not possible. In physical sciences, Monte Carlo programs are now routinely used in such simulations. This is done both before the experiment is done, to ensure that the experimental parameters are chosen so that the experiment has a possibility of detecting the effect under study, and after the experiment is performed the data are compared with results of Monte Carlo simulations, one of which includes the effect be/she has observed and the other deliberately excludes it.

We have used, and are now using, Monte Carlo simulations in toxicology, for modeling the results of the rodent bioassays of the National Toxicology Program (NTP). We are studying how the choice of a significance value (P value) in the rodent bioassay affects the number of claimed carcinogens and anticarcinogens (Linkov, Shlyakhter, Li Ping, Wilson, and Gray, 1996). In an earlier paper we demonstrated how the constraints that are set by limited sensitivity, and the observed relationship between toxicity and carcinogenicity affect correlations of carcinogenic potency (Shlyakhter, Goodman, and Wilson, 1992). This simulation involved many steps:

(a) simulating the distribution of toxicity of the chemicals in the rodent bioassays,

(b) simulating the approximate carcinogenicity/toxicity relationship including using a Monte Carlo program to include the random fluctuations in this relationship,

(c) simulating the rodent bioassay experiment, including the random sampling error,

(d) analyzing the results of that simulated experiment for carcinogenicity, and then

(e) studying the correlations between the "carcinogenic potency" calculated from the simulated results.

Attempts to discuss this analytically were less complete, and resulted in confusion and criticism. We present this as an example of how a complex simulation, we still believe the most complex in its field, can be performed readily by Monte Carlo methods.

The purpose of this paper is to illustrate how these Monte Carlo techniques might be helpful in studying potential biases, and hence uncertainties, in epidemiological studies. These biases are becoming steadily more important as epidemiologists search for smaller and smaller effects. Yet epidemiologists still routinely quote only the statistical sampling error for their studies. Such biases can be viewed as analogs of systematic uncertainties in physical measurements. The aim of many modern epidemiologists is to quantify these biases and uncertainties as much as possible. We suggest that Monte Carlo simulations are a useful tool to this end. The idea of using Monte Carlo studies in epidemiological design is not new. Feinstein and Landis (1976) used them to quantify the effect of residual biases on the frequency of spuriously significant results in randomized clinical trials. More recently, Richardson and Gilks (1993) demonstrated how Bayesian estimation of risk factors in the presence of measurement errors can be carried out for common epidemiologic designs. In this they used Gibbs sampling, an advanced sampling technique that is gaining increasing popularity in statistical and biomedical applications (Gilks *et al.*, 1993). These are approaches that are complex enough that they might inhibit widespread application. To illustrate this, we give two examples of using simple Monte Carlo simulations. In the first, we show how the effects of exposure misclassification can create false positives in an epidemiological study, and in the second, an effect we call "binning bias," we show how big a bias can be created by choosing the boundaries of the bins in which the data are presented and calculated *after* the data have been collected.

SIMULATION OF EXPOSURE MISCLASSIFICATION IN CASE-CONTROL STUDIES

We simulated a case control study with a population exposed to a *nonharmful* agent. We assume that everyone has a small probability, p=0.001, of contracting a certain disease, and study the statistical association of the incidence of this disease with exposure to the agent in a 2 X 2 contingency table (Table 1). As is usual for epidemiological studies, the results are presented in terms of a Rate Ratio (*RR*). The uncertainty in sampling makes the *RR* uncertain, and for large samples, the estimate of *ln(RR)* follows a normal (Gaussian) distribution (Rothman, 1986). Epidemiological studies present an Odds Ratio (OR) which is an unbiased estimator of the Risk Ratio. Uncertainty in the results of epidemiologic studies is reported as 95% confidence intervals (95% CI) for the Odds Ratio, *OR*, which represent uncertainty in the value of *RR*. If the agent is *nonharmful*, the odds ratio should be close to unity, and deviate from it only by statistical, sampling error. Therefore we define the "true" odds ratio to be unity (OR=1) (odds ratio is the

Table I. Contingency Table for Postulated Study

	Cases	Controls
Exposed	a	b
Unexposed	c	d

The Odds Ratio, (OR)=ad/bc
Simulation with exposure misclassification:

	Cases	Controls
Exposed	$a_1=a+f_1c$	$b_1=b+f_2d$
Unexposed	$c_1=c(1-f_1)$	$d_1=d(1-f_2)$

measure of relative risk in case-control studies). We assumed that in a small fraction (f) of subjects the exposure status was misclassified, and we further assumed that this misclassification can depend upon whether the subject is a "case" or a "control." The question asked was how often do the 95% confidence intervals cover this true value under different assumptions about the fraction (f) of subjects for whom exposure status has been misclassified? The simulation was repeated 1,000 times, each with a different set of random numbers, to reduce the statistical (sampling) error in the answers.

A population sample was considered with 100,000 exposed and 100,000 subjects. The number of "control" subjects was chosen to keep, on average, the numbers of "cases" and "controls" equal. In this example the expectation values for cases was 100, so that the number of exposed plus nonexposed controls was assumed to be 200. For each "case-control simulation," the computer simulated the numbers in four cells of the 2x2 Table 1: exposed cases, *a*, exposed controls, *b*, nonexposed cases, *c*, and nonexposed controls, *d*. In the simulation, the *observed* numbers will differ from the expectation by the sampling error.

The effects of misclassification were simulated by moving a random fraction of subjects across the cells of the 2x2 table. It was assumed that the *observed* numbers in each cell are a_1, b_1, c_1, d_1. Here $a_1=a+f_1 \cdot c$, $c_1=c(1-f_1)$, $b_1=b+f_2d$, $d_1=d(1-f_2)$, and f_1, f_2 are random numbers representing the fractions of misclassified subjects among cases and controls (see Table 1).

Specifically, we assumed that f_1 followed a normal distribution with zero mean and standard deviation σ truncated at zero so that only positive values of f_1 were allowed.

$$P(f_1) = \frac{1}{\sqrt{2\pi}} e^{\frac{-f_1^2}{2\sigma^2}} \qquad f_1 > 0$$

$$P(f_1) = 0 \qquad f_1 > 0$$

The parameter σ sets the scale of the misclassification rate.

The value of f_1 is randomly determined from this distribution in each Monte Carlo simulation. This describes a situation where all truly exposed cases were classified correctly but some nonexposed cases were classified as exposed. For f_2 a nontruncated normal distribution with zero mean and the same standard deviation σ was assumed.

$$P(f_2) = \frac{1}{\sqrt{2\pi}} e^{\frac{-f_2^2}{2\sigma^2}} \quad \text{for all } f_2$$

This simulates a situation where exposed and nonexposed controls were equally likely to give wrong answers about their exposure history.

This example was deliberately chosen to simulate a tendency (which has often been suggested in epidemiological studies) of cases to recall better (and sometimes to exaggerate) their exposure history as compared with controls. For each of the 1,000 different simulations (in a given set of simulations), different misclassification fractions f_1 and f_2 are randomly chosen. The *rate* of misclassification is determined by the parameter σ which remains the same for each of the 1,000 simulations.

If one sets $f_1 \equiv f_2 \equiv f$, one can simulate a study with a fixed amount of misclassification. However, in many cases we only have a probabilistic estimate of the misclassification to use a distribution for f_1. If one sets f_2 to be larger than f_1, or the distribution happened to be larger, one simulates a differential misclassification. We deliberately choose a *distribution* for f_1 and f_1 and f_2 to simulate either of a group of studies with different misclassification rates or a single study where the actual misclassification rate is unknown.

The simulated odds ratio was calculated to be $OR = a_1d_1/b_1c_1$; the upper and lower bounds of the 95% Confidence Interval (CI) were calculated to be: OR

$$exp(\pm 1.96(1/a_1+1/b_1+1/c_1+1/d_1)^{1/2}).$$

We illustrate these results by plotting a cumulative distribution of the normalized deviation x from the "expectation" value (unity) of the Odds Ratio (OR). This procedure is similar to that used by one of us in a graphical representation of unsuspected (systematic) errors (Shlyakhter, 1994). The normalized deviation x is given by $ln(OR)/SE(ln(OR))$ where SE is the standard error. Accordingly x was calculated to be $x=ln(OR)/(1/a_1+1/b_1+1/c_1+1/d_1)^{1/2}$ and the cumulative distribution of the x values in the 1,000 simulated studies is plotted in Figure 1. In this plot the intersection with the vertical at the deviation value of 1.96 (the value of y with x = 1.96) is the "fraction of false positives" at this chosen "level of significance" given by x. The intersection with the horizontal at the cumulative probability (fraction of false positives) = 0.025 is the normalized 95% confidence bound. For σ=0.01 the misclassification is minuscule, and the results of Monte Carlo calculation follows a Gaussian distribution - as it should (verifying that our calculation is correct). But even for a relatively small fraction of misclassification (σ=0.05), the tails extend far beyond the Gaussian distribution and the upper 95% confidence bound (corresponding to 0.025 on the y axis) is given by x = 3.2 instead of x = 1.96 for the Gaussian distribution. Looked at another way, the probability that errors exceed 3.8 standard deviations is 2.5% compared to the much smaller 1/1,000 for the Gaussian distribution.

Figure 1. The results of Monte Carlo simulations illustrating the effect of exposure misclassification on the frequency of false-positive results in case-control studies. Cumulative probability that apparent normalized logarithm of the odds ratio, (OR), x=ln(OR)/SE(ln(OR)) exceeds given value is shown for several values of the parameter σ. This parameter represents the average fraction of subjects with misclassified exposure status (see article text for details). The true value of OR=1 (the risk is not elevated). Studies that produce values x>2 are false-positive because the lower bound of the 95% confidence interval (95%CI) for OR is above the true value OR=1.

Differential misclassification of exposure status produces both false-positive findings and false-negative findings, but the increase of false negatives is not nearly as much as the increase of false-positives. This is illustrated in Figure 2 where both Cumulative Distribution Functions (CDFs) and complementary Cumulative Distribution Functions (1-CDFs) are shown with (σ=0.1) and without (σ=0.0) misclassification. As before, the true *OR=1*. Statistically significant false-negative findings ("exposure is beneficial") occur when the upper bound of the 95% CI is below one; this corresponds to x more negative than 1.96.

ANALYSIS OF POOLED STUDIES

As epidemiologists look at smaller and smaller effects, no one epidemiological study is large enough to provide a statistically significant result. For example, the various studies of workers at nuclear power plants have been pooled together to provide a "pooled" result. If each of the studies was free of biases and other nonstatistical errors, the combination is straightforward.

The summary estimate of *OR* from n studies pooled together is calculated as follows. First, we assign to each study (i) a weight, $w_i=1/var[ln(OR_i)]$, where *var* is the variance of the lognormal distribution; w_i is inverse of the squared width of the 95% CI on the log scale. These weights are used in calculation of the summary odds ratio *OR* and 95% *CI*: $ln(OR)=(\sum w_i ln(OR_i))/\sum w_i$, $1/var(ln(OR))=\sum(1/var(ln(OR_i))$ (Greenland, 1987). But this does not tell us what to do about various biases and the nonstatistical errors.

Monte Carlo simulations can also help in the understanding of how uncertainties change when several studies are pooled together. As before, we simulate a set of studies conducted on the population exposed to a *nonharmful* agent so that the "true" odds ratio *OR=1*. We ask the following question: for a given fraction of subjects with misclassified exposure status, how does pooling several studies together affect the probability that the 95% confidence intervals cover this true value?

We again consider a population sample with 100,000 exposed and 100,000 nonexposed subjects and assume that both groups have the small probability, p=0.01, (the same for both exposed and nonexposed groups, but a larger probability than in the previous set of simulations) of contracting the disease under study. For the *i-th* "case-control study," we simulate the "true" numbers of exposed cases, a_i, exposed controls, b_i, nonexposed cases, c_i, and nonexposed controls, d_i and the "apparent" numbers, a_{1i}, b_{1i}, c_{1i}, d_{1i}; the difference between "true" and "apparent" numbers accounts for exposure misclassification. For each study, we then calculate the simulated odds ratio, $OR_i=a_{1i}d_{1i}/b_{1i}c_{1i}$, the upper and lower bounds of 95% CI:

$$OR_i \cdot exp(\pm 1.96(1/a_{1i}+1/b_{1i}+1/c_{1i}+1/d_{1i})^{1/2}).$$

We again calculate the normalized deviation from the null value, $x=ln(OR)/SE[ln(OR)]$, and plot the cumulative distribution of x values. A set of pooled studies will produce false-positive results if the apparent value of pooled estimate $ln(OR)$ is more than its two standard errors away from the null value. Results

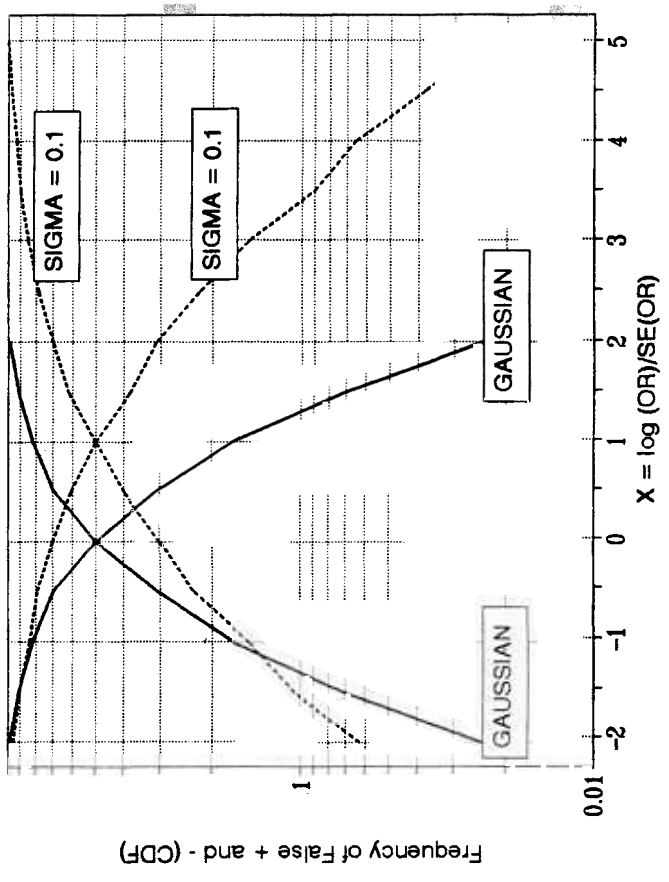

Figure 2. The probability of false-positive (x>2) and false-negative (x<-2) findings for true OR=1.00. For σ=0.01, probabilities of false-positive and false-negative findings are both close to 2.5% as they should be. For σ=0.1, the fraction of false-negatives is 6%, while the fraction of false positives is 30% due to differential misclassification of exposure status.

of 1,000 trials for individual studies (n=1) and combinations of n=5, n=10, and n=30 studies assuming 10% average misclassification rate (σ=0.05) are presented in Figure 3. The probability of a statistically significant false positive finding, $x > 1.96$, increases from 12% for n=1 to 28% for n=5, to 40% for n=10, and to 70% for n=30. This is not an unexpected result. As the number of people in the total of all the pooled studies increases, the statistical sampling error decreases. But the biases and misclassification errors remain the same and can now dominate the total uncertainty.

The Monte Carlo technique lends itself easily to complex extensions. One can simulate, for example, the pooling of a set of studies where the average rate of exposure misclassification (determined by σ) varies from study to study in a defined way.

ADJUSTMENT OF PARAMETERS AFTER THE DATA COLLECTION

It is a well-known principle of statistics that in assessing the statistical probability of an event or experimental result, decisions about what data to include and exclude, the size of the cohort and the boundaries of the bins in which the data are collected must not be modified to maximize the result. Feynman had a dramatic way of expressing this to his freshman class. Coming into class he said, "You know, the most amazing thing happened to me tonight. I was coming here, on the way to the lecture, and I came in through the parking lot. And you won't believe what happened. I saw a car with the license plate ARW 357! Can you imagine? Of all the millions of license plates in the state, what was the chance that I would see that particular one tonight?" (Goodstein, 1989) We can easily work it out: 3 is one out of 10 numbers, 5 is one out of 10 numbers, 7 is one of 10 numbers, A is one of 26 letters, R is one out of 26 letters, and W is one out of 26 letters. If we multiply these numbers together we find a low probability of one in eighteen million. Yet Feynman saw it. This commonplace experience does not seem that improbable. What is the answer to this paradox?

As presented, the answer to this paradox is obvious: Feynman did not ask the question about the particular license plate until he knew the answer. It then made no sense to ask the question.

This appears in disguised forms in many branches of science. When a scientist makes this mistake we have called the error "falling into the Feynman trap" (Shihab-Eldin, Shlyakhter, and Wilson, 1992). When this leads to an error in an experimental science, the experiment can be repeated and the mistake corrected. In the field of epidemiology, there are several well-known examples of such errors persisting into public policy. In addition to the radiation examples in the paper noted above, we note that in an early study of whether soft tissue sarcomas are caused by herbicide sprays (2-4-5T), the paper included in the analysis those cases that had brought the subject to the authors' attention in the first place (Eriksson *et al.*, 1981). This alters the statistical validity and makes it only a "hypothesis generating" study. It is therefore especially important for any future investigator to be sure that no such logical errors are made.

Ideally, in epidemiology, a prospective study is performed and a rigid protocol for the study is developed before any cases even develop. But few scientific experiments and few epidemiological studies keep to this ideal. It is therefore important to

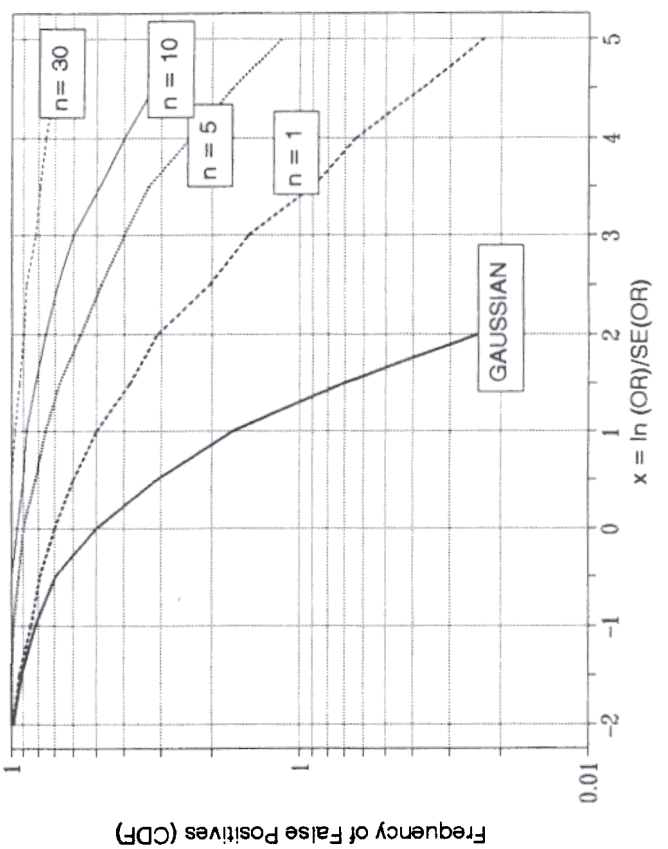

Figure 3. The probability of false-positive findings in case-control studies for individual studies (n=1) and combinations of n=5, n=10, and n=30 studies assuming that exposure status was differentially misclassified for 5% of subjects (σ=0.05).

understand the extent of any possible bias or uncertainty that can be introduced by such a failure. In this study we analyze one such possible failure: the adjustment of the boundaries of the bins into which the data are sorted and analyzed *after* the data have been collected and seen by the investigator. That this can lead to error (a version of the Feynman trap) is known to most epidemiologists. The present authors have discussed this with a small (nonrandom) sample of epidemiologists who regard such fine tuning of bin boundaries as the usual practice. One good epidemiologist stated to us that he routinely varies the bin boundaries to ensure that the data are "robust" with respect to such a change, and if they seem to be, then he chooses the boundaries that make them look the best in the final publication. Ensuring robustness in this way can avoid the most egregious of errors, but a more structured approach would be preferable.

In this simulation, we investigate the magnitude of the errors that arise from incorrectly adjusting bin boundaries in full knowledge of the data to maximize the effect. In situations where the error introduced is small, no problems arise. In situations where the error introduced is large, the level of statistical significance can be appropriately modified. It will be shown that in some realistic circumstances the errors introduced are large. This being so, we urge epidemiologists to state in their papers the extent to which the protocols were rigidly chosen in advance.

Although we have chosen this particular set of simulations with a particular controversial epidemiological study in mind, where it has been informally suggested that the authors deliberately did so adjust the bins, we feel to make a specific reference would detract from the general argument that we are endeavoring to discuss and merely turn the paper into an undesirable *ad hominem* discussion.

SIMULATION OF THE BINNING BIAS

We assume that there is an epidemiological study where data are collected into six bins. For example, they can be six different exposure categories. The Odds Ratio is calculated for the data in each bin. The epidemiologist expresses the result in terms of the odds ratio and the confidence limits. According to the usual rules, the confidence limits are chosen so that 95% of the data lie within the limits: 2 1/2% lie above the upper limit, and 2 1/2% lie below the lower limit. The epidemiologist then describes a situation where the lower confidence limit is greater than 1 as statistically significant.

In this example, 2.5% of all studies where the true odds ratio is unity will be described as statistically significant. The first (very elementary) example in which an unwary epidemiologist might modify the way of presenting the results after the data were seen is to decide to choose only the one bin out of six where the odds ratio deviates most from unity. Then the number of false positives becomes 6 X 2.5% = 15% instead of the usual 2.5%. This is an example of an error emphasized by Tippett, which elsewhere we have called the "Tippett trap" (Shihab-Eldin *et al.*, 1992). This typically occurs when a distribution of cases is plotted on a map grid, each element of the map grid being a "bin" for this purpose. If the distribution of people in the grid is unequal, this simple calculation is no longer applicable, but the true calculation can be made by Monte Carlo methods. But there are more sophisticated errors.

In the more sophisticated situation, we consider a case-control study with N cases (where N is varied from 50 to 2,000) and an equal number of controls, taken from a large pool of 10,000 persons. In the simulation, each case and each control has a dose value assigned randomly but with equal probability of lying between 0 and 1. One such simulation is shown in Figure 4. Figure 4B shows the distribution of simulated cases with dose, and Figure 4C shows the distribution of simulated controls with dose. "Exposed" is inexactly defined as a dose between 0.5 and 1; and "unexposed" is defined as a dose between 0 and 0.5. We then calculated a simulated Odd Ratio between exposed and unexposed. We normalize this Odds Ratio as before by calculating the parameter x = [ln (OR)]/[SE (OR)]. We see that for this case x = 1.5, which can occur by chance 8% of the time and is therefore not statistically significant. Now we consider the situation where a hypothetical analyst, seeing the value of x = 1.5 suspects that a real dose related effect exists. In this example the hypothetical analyst decides to see how the Odds ratio changes as he adjusts the bin boundaries. He moves the boundary progressively from 0.5 - a to 0.5 + a and recalculates the Odds Ratio progressively. The numbers for a = 0.25 are shown in Figure 4A. He notices that x increases to a high value of 2.2 at a= 0.15, and *chooses* this bin boundary for data presentation.

This simulation is repeated 30,000 times. The cumulative distribution of x is plotted in Figure 5 for several values of a - the maximum range of the hypothetical analyst's search. For a = 0.01, the distribution of x is very close to Gaussian, as it should be if the simulation is done correctly. The fractional number of times x exceeds 1.96 (often called the fraction of false positives) is the y value where the distribution curve crosses the line at x = 1.96 and is 2.5%. In Figure 6 we plot this fraction of false positives against a. We see that for our assumed case of a = 0.25, the fraction of false positives increases to 7%.

We can also look at the result in a different way. We can demand, for good and sufficient reasons, that we will accept 2.5% of false positives, but no more. How much do we have to alter the significance value to ensure that this is the case? In Figure 7 the intersection of the horizontal line at 2.5% shows the confidence limit properly adjusted for the bias (or error) introduced by bin boundary readjustment. It can be seen that this readjustment increases the uncertainty, or "error,"* from 1.96 standard deviations (σ) to 2.4 standard deviations as a increases from 0 to 0.45.

This increase in false positives at the 1.96σ limit, is slightly larger when the number of cases increases but this increase saturates. The increase in the confidence limits can approximately be expressed by multiplying the statistical uncertainty by a factor (2.4/1.96)=1.22 which is approximately independent of the number of cases.

This also can be used to simulate a more general distribution of doses which is not uniform between 0 and 1. We can define a function f(d) which *is* uniform between 0 and 1 and then the argument above works for the "unexposed" where the dose lies between 0 to d_1 where $f(d_1) = 1/2$ and the "exposed" between $d_{(2)}$ were $F(d_2) = 1$.

* We note here that physical scientists often use the word "error" in statistical calculations. In any field close to medicine "error" has a pejorative meaning, often with legal consequences, and the word "uncertainty" is preferred.

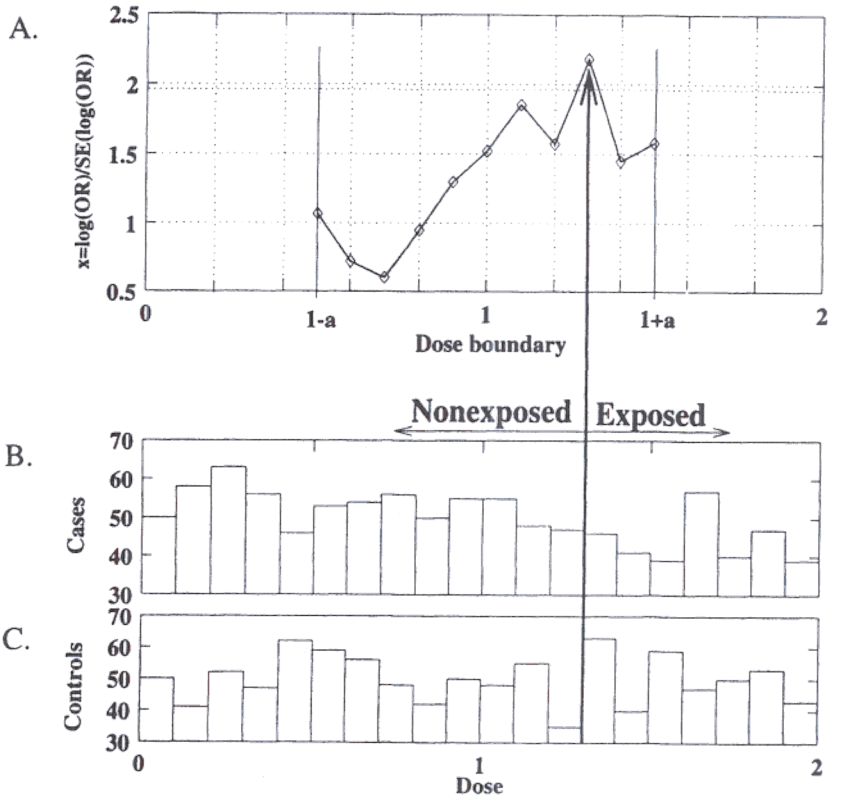

Figure 4. Illustration of the selection of the bin boundary based upon the odds ratio. In this particular simulation the bin boundary is varied from 0.25 to 0.75 (a = 0.25). The number of simulated cases and controls for each dose is shown in 4B and 4C with the calculated odds ratio in 4A. The hypothetical investigator chooses the value of a = 0.15 because that gives the largest Odds Ratio (2.2).

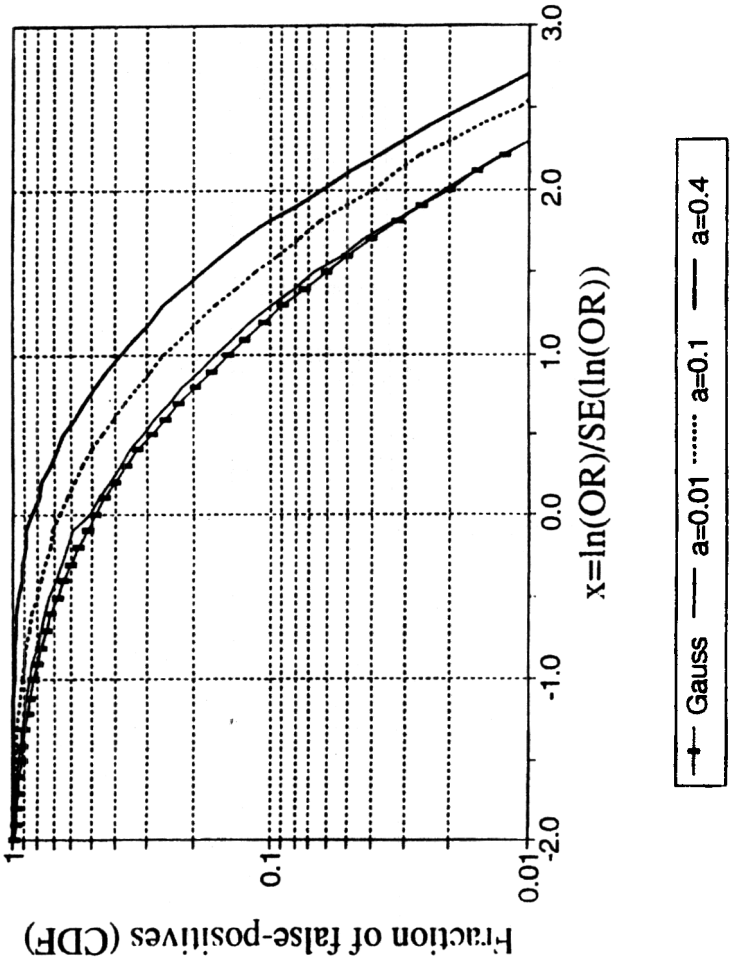

Figure 5. The cumulative distribution of the x values obtained by the arbitrary binning procedure. Distributions were obtained for different values of maximum amplitudes of bin shifting: (a=0.01, 0.1 and 0.4).

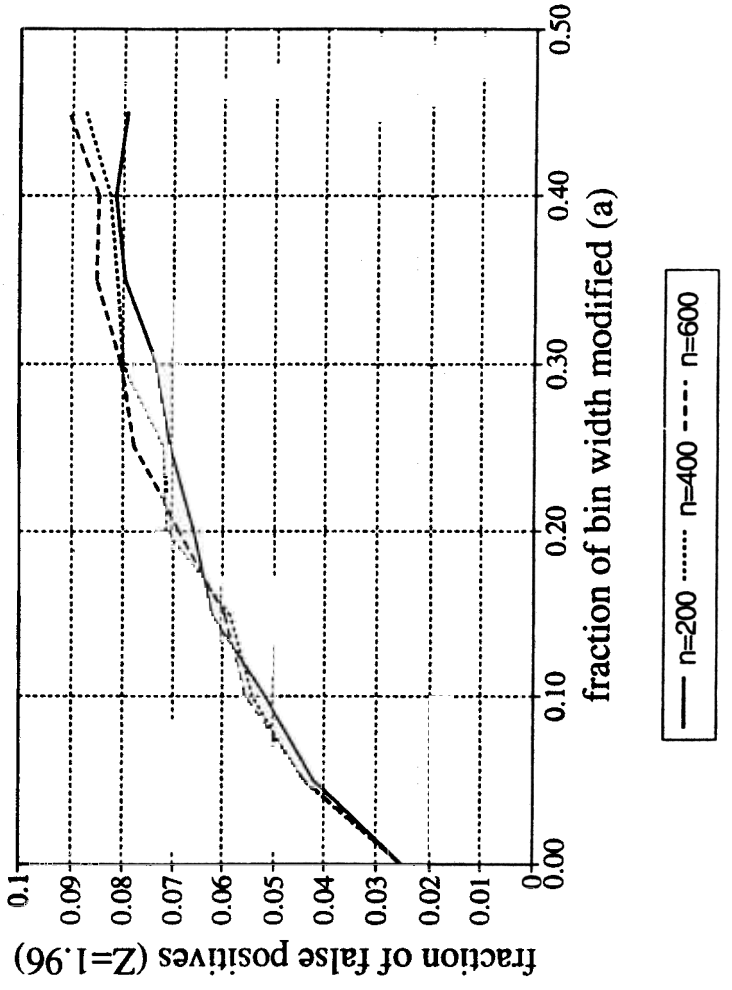

Figure 6. The fraction of false positives obtained by binning procedure as a function of the amplitude of bin shifting for different study sizes (number of cases = 200, 400, and 600).

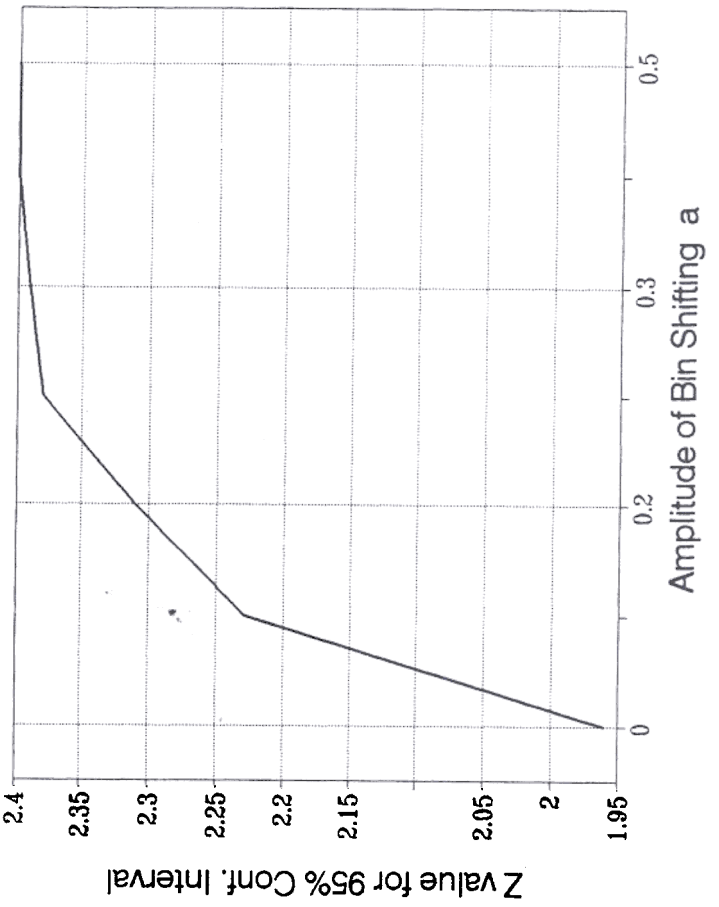

Figure 7. The 95% confidence interval for the distributions of x values obtained by binning procedure as a function of the maximum amplitude of bin shifting, a. The confidence interval was estimated as value of x which has 2.5% of cases with x above it. The sampling error of this estimate is about 0.05.

ALTERNATE PROCEDURES

Although it is possible to estimate the magnitude of a bias or error introduced by varying the bins after seeing the data, and even expand the uncertainty range to include this possibility, it would obviously be preferable to avoid having to make any such correction. Monte Carlo simulation of an experiment or study enables one to adjust bin boundaries to maximize or minimize the effect under study *before* the data are analyzed. The procedure, already widely used in the physical sciences, is to simulate the epidemiological study in advance, and to simulate any possible change in the binning or other criterion without reference to the data.

It has already become standard practice to ask questions in advance: What are the questions that we hope that this study will answer? Given the known biases and known statistical (sampling) error, can the proposed study, with its chosen protocol answer them?

These questions are usually answered by an analytic calculation. But they can also be answered by a Monte Carlo simulation, at which time these more sophisticated questions can also be asked. For example if one has a reasonable number for the expected incidence of the disease being studied, and a knowledge of the number of people and their age distribution in the study, one can decide on the appropriate bins in which to present the data using the simulated data. Then the bins so chosen can be used for presentation of the actual data. This would avoid completely the troublesome bias.

CONCLUSIONS

The interpretation of the results of observational studies and their use in regulatory risk assessment becomes progressively more difficult as epidemiologists deal with smaller risk ratios (Taubes, 1995). Usually, the reported 95% confidence intervals in epidemiological studies reflect only sampling errors and do not formally include uncertainty caused by misclassification and confounding. This makes it hard to describe an overall uncertainty in the epidemiological result. We have applied the techniques of Monte Carlo simulation to the analysis of the effects of two possible sources of systematic uncertainties in case-control studies with slightly elevated risk ratios. We show that even a small fraction of subjects with misclassified exposure status (differential among cases and controls) can cause a considerable fraction of statistically significant, but false positive, results. The effect of differential misclassification is more important for large studies where random errors are relatively small and when several studies are pooled: upon pooling, the statistical uncertainty is reduced but the misclassification uncertainty stays approximately constant.

We have seen that in an example which is not implausible, adjustment of the bins, or other choice of data presentation *after* the data are collected can result in considerable error, which can be allowed for by increasing the statistical error considerably. In high energy physics, a field in which one of us has had considerable experience, it has been noted that good experimenters emphasize in their papers any important feature of experimental design, or data analysis. If an important feature is not emphasized in a paper, the usual reason is that it has been ignored. By analogy

we suggest that it is important for each epidemiologist to make clear and obvious the extent to which the prechosen protocol was strictly obeyed in his study, and if it was not, the estimated effect of this failure. This is especially important now that epidemiologists are looking at smaller and smaller risk effects. It has, for example, been noted that if the risk ratios are less than about 2, information from disciplines other than from the epidemiological studies is essential before any association can be accepted as causal or even real. Then it is important that the scientists in the other disciplines fully understand the epidemiological studies with all their strengths and weaknesses.

Acknowledgments

We would like to thank Professor J.D. Graham and Dr. D.E. Burmaster for their interest and support.

References

Crouch, E.A.C. and Wilson R. 1981. Regulation of carcinogens. *Risk Anal.* **1**, 47.

Crouch, E.A.C., Wilson R., and Zeise L. 1983. The risk of drinking water. *Water Resour. Res.* **19**, 1359-1375.

Eriksson, M., Hardell, L., Berg, N.O., Moller T., and Axelson, O. 1981. Soft-tissue sarcomas and exposure to chemical substances: A case referent study. *Brit. J. Ind. Med.* **27**, 33.

Feinstein, A.R. 1988. Scientific standards in epidemiologic studies of the menace of daily life. *Science* **242**, 1257-1263.

Feinstein, A.R. and Landis, J.R. 1976. The role of prognostic stratification in preventing the bias permitted by random allocation of treatment. *J. Chron. Dis.* **29**, 277-284.

Gilks, W.R., Clayton, D.G., Spiegelhalter, D.J. *et al.* 1993. Modelling complexity: Applications of Gibbs sampling in medicine. *J. Royal Statist. Soc.* **B55**, 39-52.

Goodstein, D.L. 1989. Richard P. Feynman, teacher. *Phys. Today* February, 70-75.

Greenland, S. 1987. Quantitative methods in the review of epidemiologic literature. *Epidemiologic Rev.* **9**, 1-30.

Gregorio, D.I., Marshall, J.R., and Zielezny, M. 1985. Fluctuations in odds ratios due to variance differences in case-control studies. *Am.J. Epidemiol.* **121**, 767-774.

Linkov I., Shlyakhter, I., Li Ping, Wilson R., and Gray G. 1996. "Anti-Carcinogenic Responses in Rodent Cancer Bioassays," paper in preparation.

Mayes, L.C., Horwitz, R.I., and Feinstein, A.R. 1988. A collection of 56 topics with contradictory results in case-control research. *Int. J. Epidemiol.* **17**, 680-685.

Richardson, G.M. and Gilks, W.R. 1993. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *Am. J. Epidemiol.* **138**, 430-442.

Rossi H.H. and Greisen K. 1947. Cosmic ray theory. *Revs Mod. Phys.* **13**, 263.

Rothman, K. 1986. *Modern Epidemiology*, Boston, Little, Brown & Co.

Shihab-Eldin, A., Shlyakhter, A.I., and Wilson, R. 1992. Is there a large risk of radiation? A critical review of pessimistic claims. *Environ. International*, **18**, 117-151 (expanded version: Argonne National Laboratory Report, ANL-92/23).

Shlyakhter, A.I., Goodman, G., and Wilson, R. 1992. Monte Carlo simulation of rodent carcinogenicity bioassays. *Risk Anal.* **12**, 73-82.

Shlyakhter A.I. 1994. An improved framework for uncertainty analysis: Accounting for unsuspected errors. *Risk Anal.* **14**, 441-447.

Shlyakhter, A.I. and Wilson, R. 1995. Monte Carlo simulation of uncertainties in epidemiological studies: An example of false-positive findings due to misclassification. *Proc. ISUMA'95 The Third International Symposium on Uncertainty Modeling and Analysis*, pp. 685-689. University of Maryland, College Park, Maryland, September 17-20, 1995. Los Alamitos, CA, IEEE Computer Society Press.

Taubes G. 1995. Epidemiology faces its limits. *Science* **269**, 164-169.

Wilson, R. 1949. Random selectors for ESP experiments. *J. Soc. Psych. Res.* **48**, 213.

Wilson, R.R. 1950. Monte Carlo calculation of lead in showers. *Phys. Rev.* **79**, 204.

Wilson, R.R. 1951. The range and straggling of high energy electrons. *Phys Rev.* **4**, 100.