*Statistical Science*
2003, Vol. 18, No. 4, 466–480

# Epidemiologically Based Environmental Risk Assessment

**Louise Ryan**

*Abstract.* Environmental health research aims to discover and understand the links between environmental exposure and disease and to inform the regulatory community so that society can be protected against cancer, birth defects and other adverse health effects associated with chemical, industrial and other exposures. Statistical science has a critical role to play in terms of providing the appropriate tools to design and analyze the studies needed to address the questions of interest, as well as quantifying risks and characterizing uncertainty. Recent years have seen some dramatic changes in the way that environmental risk assessment is accomplished. One such change is a move away from a traditional reliance on toxicological studies in animals to incorporate more epidemiological data. This shift has been facilitated by scientific advances that now allow researchers to accurately characterize human exposures in a variety of settings, as well as to measure genetic and other biomarkers that reflect subtle health effects and variations in susceptibility. This article will use a high profile case study to highlight some of the challenging statistical issues arising from this shifting emphasis from animal based toxicology to environmental epidemiology in the risk assessment world. Among the topics to be discussed are the uses of biologically based models and biomarkers, as well as the role of Bayesian methods to characterize uncertainty due to population heterogeneity, unmeasured confounders, exposure measurement error and model uncertainty.

*Key words and phrases:* Quantitative risk assessment, arsenic, carcinogenicity, dose response.

## 1. INTRODUCTION

It has long been known that environmental exposures can adversely affect human health. Attempts to regulate dangerous exposures also have a long history (Moeller, 1992). Much of modern day environmental health research has its roots in the post-industrial revolution years when increasing urbanization led naturally to concerns about the safety of food, housing, sanitation, industrial waste and other aspects of public works that influence human health. As progress was made on more basic issues, the field evolved to encompass more subtle concerns, especially the effects of chemical pollutants on cancer, respiratory and heart disease. This interest in characterizing more subtle effects led to the need for sophisticated statistical and mathematical methods to quantify risks and to help inform decision makers and regulators charged with setting environmental standards.

During the latter part of the late 20th century, much of environmental health research and regulation was based on toxicological studies in rodents. By conducting controlled experiments in genetically homogeneous animal populations, investigators could control extraneous sources of variability and also boost study power by using higher exposure levels. Also, conducting studies in animals allows for a

*Louise Ryan is Professor, Department of Biostatistics, Harvard School of Public Health, Harvard University and Dana-Farber Cancer Institute, 4th Floor, Room 409, 655 Huntington Avenue, Boston, Massachusetts 02115 (e-mail: lryan@hsph.harvard.edu).*

toxicological assessment prior to widespread exposure in humans (this argument is particularly important for the pharmaceutical industry). While toxicological studies in animals are likely to continue as the "backbone of most risk assessment determinations" for the foreseeable future (Olden and Guthrie, 1996), recent advances in the field of environmental epidemiology have led to an increasing reliance on human data.

The purpose of this paper is to highlight some of the challenging statistical problems motivated by modern environmental health risk assessment involving epidemiological studies in humans. After a brief history of risk assessment in the United States, a case study will be presented and used to highlight weaknesses in the traditional paradigms. We will discuss several areas which are ripe for the development of new methods, and we conclude with a general discussion.

## 2. ENVIRONMENTAL HEALTH RISK ASSESSMENT IN THE UNITED STATES—A BRIEF HISTORY

Effective regulation of toxic substances and exposures requires a careful balance of scientific knowledge and uncertainty, economic and political considerations. Most countries have their own regulatory system, though they are strongly influenced by recommendations from international bodies such as the World Health Organization (WHO) and International Agency for Research on Cancer (IARC). This section provides some brief background on current regulatory practice in the United States.

A number of different U.S. federal agencies have authority to set and enforce laws to protect the public against dangerous environmental exposures. The Environmental Protection Agency (EPA) has broad-reaching authority over water, air and land quality standards, while other agencies such as the Occupational Safety and Health Administration (OSHA) and the Nuclear Regulatory Commission play more specialized roles. One of the oldest regulatory agencies, the Food and Drug Administration (FDA), dates back to the late 1800s and is charged with ensuring not only the efficacy, but also the safety of drugs and medical devices. While the details of regulatory practice and terminology vary from agency to agency, there are important common themes. These days, most agencies follow the broad guidelines set out in a National Academy of Sciences publication perceived by many as the "risk assessment bible" [National Research Council (NRC), 1983]. This report distinguishes risk assessment from risk management, the former using science to define and quantify health effects associated with environmental exposures, and the latter weighing those findings with social, economic and political concerns to set regulatory standards. Although sound statistical thinking has a role to play in all stages of the risk assessment process, statisticians have been most active in the areas of hazard identification (which involves deciding whether or not an exposure is causally associated with an adverse health effect, and hence relies heavily on hypothesis testing) and dose response assessment (which involves quantifying the dose–response relationship once a substance has been deemed hazardous to human health).

The Food, Drug and Cosmetics Act of 1938 set the stage for much of modern risk assessment. Of various revisions and amendments to this act over the years, perhaps none has garnered more attention nor had a larger influence on the regulatory community than the infamous Delaney Clause of 1958, which posited a zero tolerance policy for suspected carcinogens, specifying that "no additive shall be claimed safe if it is found to induce cancer when ingested by man or animal."

When it was established in 1970 (see http://www.epa.gov/history/), the EPA drew heavily on many of the FDA's regulatory practices. Because outright banning of carcinogens was not always practical, the EPA's version of the Delaney Clause involved determination of a virtually safe dose (VSD), or the dose corresponding to an "acceptable cancer risk," typically $10^{-6}$. Related to such low-dose risk estimation was the so-called unit risk, which corresponded to the increase in risk of an adverse effect associated with a one-unit increase in exposure. This movement toward a quantification of environmental health risk attracted many statisticians to the field. Because a VSD could not generally be determined experimentally, the general approach was to conduct a high-dose study in laboratory animals, apply dose response modeling techniques and extrapolate. Given a dose response function $p(x)$ representing the lifetime probability of developing a tumor for an animal exposed to dose level $x$, a VSD is computed by solving $r(x) = 10^{-6}$, where $r(x)$ is a measure of excess risk, additive or multiplicative excess risk models being

$$r(x) = p(x) - p(0) \quad \text{or} \quad r(x) = \frac{p(x) - p(0)}{1 - p(0)},$$

respectively (see Gart et al., 1986, Chapter 6). Choices for $p(x)$ range from biologically motivated models, such as the multistage model (Armitage and Doll, 1954), to a variety of more empirical, statistical models (see Piegorsch and Bailer, 1997, for review; also Morgan, 1992). A lower limit on the true VSD could be obtained through use of a statistical confidence limit, for example, through use of the delta method (see Gart et al., 1986, page 270). While technical aspects of calculating a VSD were relatively straightforward, there were still a variety of interesting statistical issues to address. For example, the classic paper by Hoel and Walburg (1972) alerted statisticians to the bias that can result from ignoring age of death when computing lifetime tumor incidence rates in a rodent tumorigenicity study. Numerous papers ensued, including three-state models for carcinogenicity, following the seminal work of Kodell and Nelson (1980). While work on three-state modeling of carcinogenicity data is important and still of interest today (for some recent examples, see Dunson and Dinse, 2002; Mancuso, Ahn, Chen and Mancuso, 2002), some statisticians have focused on broader topics. Others have criticized cancer risk assessment methodologies at a fundamental level (see, e.g., Freedman and Zeisel, 1988; Lin, Gold and Freedman, 1995; also Abelson, 1995).

Because the Delaney Clause made no mention of noncancer risks such as birth defects, neurological effects and so on, regulatory methods for such endpoints developed along a different path. Based on the concept that noncancer health effects are likely to operate according to a threshold mechanism, risk assessment for noncancer endpoints has been based on determination of a "no observed adverse effect level" (NOAEL), defined as the experimental dose level immediately below the lowest dose that produces a statistically or biologically significant increase in the rate of adverse effects, compared to controls. An appropriate human exposure level is generally derived by dividing the NOAEL by a "safety factor" of usually 100 or 1,000 to allow for the possibility of sensitive subpopulations, extrapolation from animal data to human risk and other sources of uncertainty (see NRC, 1994). Various regulatory agencies use different terminology to describe these recommended regulatory levels. For example, the EPA refers to this safe daily concentration as the reference dose (RfD) while the FDA uses the term allowable daily intake (ADI).

During the 1980s and 1990s, use of NOAELs for noncancer risk assessment became controversial, in large part due to the emergence of some serious statistical flaws with the approach (see, e.g., Gaylor, 1983; Kaplan, Hoel, Portier and Hogan, 1987; Kimmel and Gaylor, 1988). For instance, because the NOAEL must correspond to one of the experimental doses, its value can vary by orders of magnitude under repeated experimentation, yet this statistical variation is ignored. Estimation of the NOAEL is anticonservative with regard to sample size: since the NOAEL is based on comparison to control levels, large studies have higher power to detect small changes and therefore produce lower NOAELs. In contrast, more variable, smaller studies tend to produce higher, less conservative NOAELs. A landmark paper by Crump (1984) proposed replacing the NOAEL by a so-called benchmark dose (BMD), based on a dose response modeling approach and defined as a lower 95% (or 99%) confidence limit on the dose corresponding to a moderate increase (e.g., 1%, 5% or 10%) over the background rate.

From a statistical perspective, computation of a benchmark dose is no different than for a VSD, except that a more modest risk level of 1%, 5% or 10% is used instead of $10^{-6}$. Several authors have argued (see, e.g, Allen, Kavlock, Kimmel and Faustman, 1994; Leisenring and Ryan, 1992) that a NOAEL from a typical-sized toxicological experiment will correspond, on average, to a dose level close to the 5% or 10% risk level. The move toward the use of a benchmark dose for noncancer risk assessment has also influenced thinking about cancer risk assessment. For example, the EPA has proposed the use of a benchmark dose approach for cancer risk assessment, unless a strong biological justification could be made for extrapolation based on a particular dose response model (U.S. EPA, 1999).

There are many topics that could be addressed in a paper on statistical methods for modern environmental health risk assessment. The chosen focus here, however, is a series of broad statistical challenges arising from the relatively recent trend toward the use of epidemiological data for quantitative risk assessment.

## 3. EPIDEMIOLOGICALLY BASED RISK ASSESSMENT

Often, good quality human dose–response data can be obtained from an occupational setting. For example, EPA based its recent risk assessment for the combustion by-product 1,3-butadiene on studies that had been conducted in occupationially exposed rubber factory workers (U.S. EPA, 2002). The National Research Council estimated high cancer risk due to radon exposure based on studies in miners (NRC, 1988). In this section, we will discuss a high-profile case study where the first evidence of adverse effects arose from an accidental poisoning. Although superficially it might be argued that risk assessment based on human data involves the same principles as that based on animal data, we will see that some unique and challenging issues arise.

### 3.1 Arsenic in Drinking Water

Establishing appropriate regulatory standards for arsenic in drinking water has been a source of considerable controversy and provides a fascinating example of the complex interplay between science and policy. Arsenic is a naturally occurring metal and had been used for medicinal purposes for well over a century before evidence began to emerge regarding serious adverse health effects associated with chronic exposures (NRC, 1999). Some of the most compelling data arose from a rural population in southwestern Taiwan who had been exposed to high levels of arsenic in drinking water after primitive "tube wells" had been sunk in a postwar effort to increase supplies of fresh drinking water in the region. Earlier reports were concerned with skin lesions, including blackfoot disease (a condition that can lead, in its most extreme form, to limb loss) and a nonlethal form of skin cancer (Tseng et al., 1968). However, more mature studies based on the same population eventually pointed to arsenic's being associated with several life-threatening conditions, including bladder and lung cancer (Chen, Chuang, Lin and Wu, 1985). Evidence of carcinogenic effects of arsenic have also emerged from studies in other parts of the world, along with the possibility that chronic arsenic exposure can also cause diabetes and cardiovascular disease (NRC, 1999).

Although the World Health Organization had long since recommended a standard of 10 parts per billion (ppb), the U.S. standard for arsenic in drinking water remained at an interim level of 50 ppb in the late 1990s, despite increasing pressure to promulgate a revision. Unable to reach a consensus on how to do so, the EPA sought advice from the National Academy of Sciences (NAS), which subsequently released a report (NRC, 1999) confirming unacceptably high risks at 50 ppb and urging the EPA to establish a lower water standard. The EPA established a new maximum contaminant level (MCL) of 10 ppb in late 2000, and it was hoped that the controversy would be ended. However, not long after the Bush presidency was established in 2001, the new standard was revoked and EPA was instructed to go back and seek a decision based on "better science." The EPA again turned to NAS, which issued an updated report on September 11, 2001, reiterating its earlier findings and pointing to studies released since the earlier report which added to the weight of evidence supporting a standard closer to 10 ppb. EPA reinstituted the revised standard of 10 ppb in November 2001.

Many of the core issues that made the arsenic story so controversial were inherently statistical in nature. Being one of very few compounds that appear to be carcinogenic in humans but not in animals, quantitative risk assessment for arsenic has had to rely exclusively on epidemiological data. A 1988 EPA report had used prevalence data (see Table 1) from the Tseng study to estimate an excess lifetime skin cancer risk of between 3 and 7 per 1,000 for a typical U.S. resident exposed over their lifetime to an arsenic level of 50 ppb (U.S. EPA, 1988). A number of issues with EPA's statistical analysis contributed to a stalemate in terms of their using the results to implement new guidelines. Concerns related to exposure assessment were foremost. First of all, the Tseng study had not been designed for a dose response analysis. Instead of assessing each individual's exposure level, the study had simply classified each subject as being exposed to low, medium or high arsenic levels, based on the levels measured in the wells from the village where they lived. This gave the study a so-called ecological design, which is generally considered to be the least desirable basis for use in risk assessment (NRC, 1991).

TABLE 1

*Data reported by Tseng et al.* (1968), *adapted from U.S. EPA* (1988, *Table B-l*); *entries show male population at risk, followed by number of skin cancer cases in parentheses*

| Arsenic concentration | Age group (in years) | | | | Total |
|---|---|---|---|---|---|
| | 0–19 | 20–39 | 40–59 | ≥ 60 | |
| Low (0–30 ppb) | 2714 (0) | 935 (1) | 653 (4) | 236 (11) | 4538 (16) |
| Medium (30–60 ppb) | 1542 (0) | 531 (2) | 371 (18) | 134 (22) | 2578 (42) |
| High (> 60 ppb) | 2351 (0) | 810 (18) | 566 (56) | 204 (52) | 3931 (126) |
| Unknown | 4933 (0) | 1699 (3) | 1188 (61) | 429 (64) | 8249 (128) |

As discussed by many authors (e.g., Wakefield, 2003; Greenland and Robins, 1994), the main concern with the use of ecological data is the potential for bias associated with the omission of important individual-level confounding factors. In the case of the Tseng study, however, it is reasonable to argue that concerns about confounding are relatively minimal: the study area was relatively homogenous from a socioeconomic standpoint, comprising a relatively stable, but poor, rural population. Of more serious concern here was the potential for measurement error to influence the results. Unfortunately, while it is easy to raise general questions and concerns, relatively little work has been done toward quantifying the bias that could be involved. Prentice and Sheppard (1995) argue that, so long as appropriate adjustments are made for confounding, ecological studies may be subject to less measurement error than studies where exposures are measured at the individual level. Finally, as seen in Table 1, exposure levels could not be determined for many subjects in the Tseng study.

In light of all the problems with the Tseng skin cancer data, the NRC committee decided to focus on data related to internal cancers. In contrast to the Tseng data, the internal cancer data (reported by Chen et al., 1985) involved a more accurate exposure assessment, in that measured arsenic levels were reported for each village in the study area. Table 2 shows a subset of the raw data (two villages only), stratified by age group. Readers interested in the full data may obtain it from *Statlib* (www.stat.cmu.edu—select "get data" and search for arsenic). The literature includes relatively little discussion about the computation of a benchmark dose based on epidemiological cohort data, the only sources being a brief mention in a book chapter on metaanalysis by Wright, Lopipero and Smith (1997) and an appendix to a National Academy report on radon (NRC, 1988, page 131). We thus describe the approach in some detail.

TABLE 2
*Male lung cancer data reported by Chen et al.* (1985); *entries show village-specific median arsenic levels, person years at risk in each age group, followed by number of cancer deaths; only two villages are included; full data available at Statlib* (*www.stat.cmu.edu*)

| Village | Arsenic conc. in ppb | Age group midpoint (years) | Person years at risk | No. lung cancer deaths |
|---|---|---|---|---|
| 1 | 10 | 22.5 | 1128 | 0 |
| 1 | 10 | 27.5 | 634 | 0 |
| 1 | 10 | 32.5 | 389 | 0 |
| 1 | 10 | 37.5 | 313 | 0 |
| 1 | 10 | 42.5 | 364 | 0 |
| 1 | 10 | 47.5 | 410 | 0 |
| 1 | 10 | 52.5 | 325 | 1 |
| 1 | 10 | 57.5 | 227 | 1 |
| 1 | 10 | 62.5 | 141 | 1 |
| 1 | 10 | 67.5 | 104 | 0 |
| 1 | 10 | 72.5 | 63 | 0 |
| 1 | 10 | 77.5 | 39 | 0 |
| 1 | 10 | 82.5 | 22 | 1 |
| 40 | 698 | 22.5 | 1085 | 0 |
| 40 | 698 | 27.5 | 617 | 0 |
| 40 | 698 | 32.5 | 390 | 0 |
| 40 | 698 | 37.5 | 361 | 0 |
| 40 | 698 | 42.5 | 395 | 0 |
| 40 | 698 | 47.5 | 339 | 0 |
| 40 | 698 | 52.5 | 337 | 1 |
| 40 | 698 | 57.5 | 275 | 1 |
| 40 | 698 | 62.5 | 195 | 1 |
| 40 | 698 | 67.5 | 167 | 0 |
| 40 | 698 | 72.5 | 102 | 1 |
| 40 | 698 | 77.5 | 37 | 2 |
| 40 | 698 | 82.5 | 10 | 0 |

Poisson modeling provides a convenient and natural framework for analyzing cancer incidence data of the form seen in Table 2 (see Breslow and Day, 1987, for further discussion). Suppose the data are divided into $n$ unique covariate combinations (in our case, age and arsenic concentration), indexed by $i$. While gender could also be considered as a covariate, we describe here the approach taken by NRC, namely reporting separate analyses for males and females. Let $d_i$ denote the number of cancer deaths among the $r_i$ person-years-at-risk in group $i$, let $x_i$ be the arsenic concentration and let $t_i$ be a suitable representative age (e.g., midpoint of the age interval) for the same group. Then the Poisson modeling approach assumes

$$(1) \qquad d_i \sim \text{Poisson}\big[r_i h^C(t_i, x_i)\big],$$

where $h^C(t, x)$ is the cause-specific hazard of dying from the cancer of interest for an individual aged $t$, exposed to arsenic at level $x$. The Poisson modelling approach derives naturally from a survival analysis framework where the hazard for death from the cancer of interest is piecewise constant on intervals that correspond to the observed age groups (see Laird and Olivier, 1981). Assumption of a multiplicative exposure effect implies

$$(2) \qquad h^C(t, x) = h^C(t, 0)g(x),$$

with $h^C(t, 0)$ reflecting the cause-specific hazard for unexposed individuals aged $t$ and $g(x)$ is interpretable as a relative risk. The latter can be modeled using an appropriate parametric form, with simple models such as $g(x) = \exp(\beta x)$ being easily fitted in standard generalized linear model (glm) software (PROC GENMOD in SAS or glm in S-PLUS), declaring *age group* as a factor variable. More complicated dose response models, and additive models,

$$h^C(t,x) = h^C(t,0) + g(x),$$

can also be used, though these may require specialized programming. A parsimonious approximation to a piecewise constant age effect is to impose a parametric form on $h^C(t,0)$, for example,

$$h^C(t,0) = \exp(\alpha_0 + \alpha_1 t + \alpha_2 t^2),$$

with the values of $t$ chosen to correspond to the midpoint of each of the observed age intervals. Such an approach might be suitable for smaller data sets. Strictly speaking, model fitting should account for the grouped nature of the data, for example, by considering an EM approach (see Brumback, Cook and Ryan, 2000). In practice, it seems unlikely that this would have a substantial impact on the estimated hazard rate, especially when balanced with the magnitudes of other uncertainties, discussed below.

Of course, there are many choices for characterizing $h^C(t,x)$ (see NRC 1999, 2001; Morales et al., 2000). EPA has often favored the multistage Weibull model, which generalizes the multistage model (Armitage and Doll, 1954) to accommodate age effects, and which corresponds to putting

$$(3) \qquad h^C(t,x) = (t - t_0)_+^k \sum_{j=1}^{J} \alpha_j x^j,$$

where $t_0, k$ and the $\alpha_j$'s are unknown parameters. The term $(t - t_0)_+^k$ denotes a truncated polynomial of the $k$th degree, which takes the value 0 if $t < t_0$ and $(t - t_0)^k$ otherwise. As in the multistage model, the parameters $\alpha_0, \dots, \alpha_J$ are generally constrained to be positive. Other options include the consideration of various dose transformations (log, square root, etc.) and flexible models for age [quadratic, spline models, etc. (see Morales et al., 2000)].

Translating the Poisson modeling results to a benchmark dose requires computing $p(x)$, the risk of dying from the cancer of interest for someone exposed over their lifetime to exposure concentration $x$:

$$p(x) = \int_0^\infty S(t,x) h^C(t,x)\, dt,$$

where $S(t,x)$ is the overall probability of surviving to age $t$ for someone exposed to arsenic level $x$, and $h^C(t,x)$ is the cause-specific hazard defined above. In the context of an animal experiment, the survivorship function $S(\cdot)$ would be estimated from the observed data, along with the cause-specific hazard function $h^C(\cdot)$ [see Finkelstein (1991) for discussion related to calculating age-adjusted lifetime tumor rates in an animal study]. In the context of an epidemiologic cohort study, however, it will generally be necessary to go to other data sources for information about $S$. Indeed, the purpose of the NRC's arsenic risk assessment was to estimate a benchmark dose for the United States, even though the data being used to quantify the dose–response relationship were based on the Taiwanese population. The National Academy used U.S. life tables and cancer mortality data to estimate both $S(t,0)$ and $h^C(t,0)$, the survivorship and cause-specific hazard functions for an unexposed subject. Because life table and cancer mortality data are reported in five-year intervals, it is useful to consider a discretized version of the formula for $p(x)$. Let $h_g^C(x)$ and $h_g^O(x)$ be the cancer-specific and overall death hazards for people exposed at arsenic level $x$ during the $g$th age interval. Similarly, let $q_g(x)$ be the conditional probability of surviving through the end of $g$th age group, given survival to the beginning of that age group, for someone exposed at level $x$. Then, the unconditional probability of surviving to the beginning of the $g$th age group: $S_g(x) = \prod_{f \le g} q_f(x)$. Then, some simple algebra (see NRC, 1988, page 131) establishes the following approximation:

$$(4) \qquad p(x) \approx \sum_{g \in \mathcal{G}} \frac{h_g^C(x)}{h_g^O(x)} S_g(x)\bigl(1 - q_g(x)\bigr),$$

where the sum is over the set of age groups $\mathcal{G}$ represented in the life tables. Under the multiplicative model (2), it follows that $h_g^O(x)$ can be written as

$$(5) \qquad h_g^O(x) = h_g^O(0) + [g(x) - 1] h_g^C(0),$$

and also

$$q_g(x) = q_g(0) e^{-5[g(x)-1]h^C(t,0)}.$$

Once $p(x)$ has been estimated, the benchmark dose simply corresponds to the value of $x$ that solves $p(x) - p(0) = q$, where $q$ is the desired risk level. A Taylor series expansion of the expression in (4) establishes that a simple approximation (see also Wright, Lopipero and Smith, 1997) to the BMD is the value of $x$ that solves $p(0)(g(x) - 1) = q$.

The approach outlined here is particularly appealing when it comes to extrapolating from one population to another. For example, the arsenic analysis dose response model was estimated using data from Taiwan, yet the objective was to estimate a benchmark dose for the U.S. population. Although it is well known that baseline cancer incidence rates vary significantly from country to country, epidemiologists have argued that relative risks associated with smoking and other environmental exposures tend to be fairly constant (Breslow and Day, 1987). This argument supports the approach taken for the arsenic risk assessment, namely using an estimated relative risk based on data from Taiwan in (5), with baseline hazards for overall and cancer-specific deaths taken from vital statistics data for the United States. A major advantage of the relative risk approach is that it can even be applied when the available data related to dose response effects come from a case–control study. For example, Ferreccio et al. (2000) report on a case–control study from Chile which examines the association between arsenic exposure and lung cancer. Figure 1 plots estimated odds ratios, along with associated confidence intervals (dotted lines), reported in Table 5 of that paper, for five groupings of arsenic exposure levels (0–10, 10–29, 30–49, 50–199 and 200–400 ppb). Also shown in the plot is a linear fit to these estimated odds ratios, with the line forced through the point $(0, 1)$ and with assumed exposure levels within each exposure grouping set at the midpoint.

Despite the seeming simplicity of these arguments from a statistical perspective, a number of factors complicated the process of using the Taiwanese data to predict risks for the U.S. population. We discuss some of these briefly.

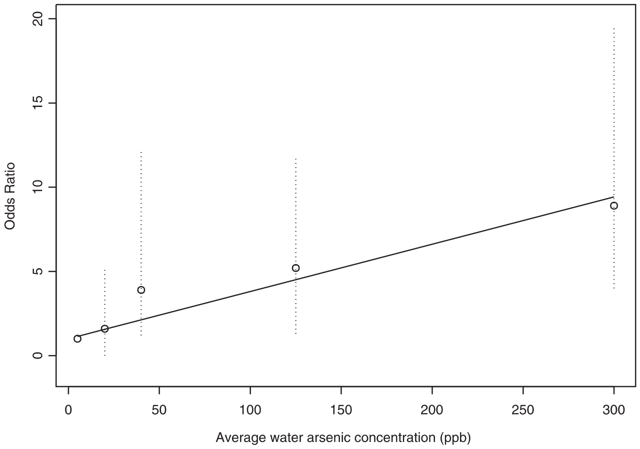

FIG. 1. *Estimated odds ratios and confidence limits from Table* 5 *in Ferreccio et al.* (2000), *along with linear fit.*

*Body weight and water consumption.* Of particular importance were assumptions related to body size and drinking water consumption in the two countries. Toxicologists commonly assume that exposure per unit body weight is the appropriate dose metric for extrapolating toxicological effects between two populations whether two animal species or two different human populations. In other words, a 100-pound person exposed to 50 micrograms of a toxicant should experience, on average, the same toxicity as a 200-pound person exposed to 100 micrograms of the same substance. In their 1988 arsenic risk assessment, the EPA argued that adjustments were needed to extrapolate results from Taiwan to the United States because (a) the current U.S. population is generally much heavier than the poor rural Taiwanese study population back in the 1960s and (b) the typical subject from the Taiwanese study was likely to drink more water per day than the typical U.S. person. In their 1988 analysis EPA assumed that a typical Taiwanese male from the study weighed 55 kg and drank 3.5 liters of water per day, while the typical female weighed 50 kg and drank 2 liters of water per day. In contrast, typical U.S. males and females were both assumed to weigh 70 kg and drink 2 liters of water per day. This means that in terms of exposure per unit body weight, the toxicity experienced by a U.S. resident drinking water contaminated with 50 ppb (micrograms per liter) of arsenic would be equivalent to a Taiwanese male drinking from a source contaminated with only $50 \times 2 \times 55/(70 \times 3.5) = 22$ ppb (36 ppb for a Taiwanese woman). To accommodate these differences, the EPA simply rescaled the exposure levels in the Taiwanese data by a factor of $50/22 = 2.27$ for males and $50/36 = 1.39$ for females, so that the resulting BMD estimate would be relevant for the United States. Clearly, variations in these assumptions could have substantial impact on estimated benchmark dose calculations! For example, assuming that a typical U.S. resident weighs 80 kg instead of 70 kg would result in a 15% reduction in the estimated BMD. In a later section of the paper we briefly discuss a more rigorous approach to addressing the uncertainty associated with variation in body weight and drinking water rates between the two populations.

*Baseline cancer rates.* Another factor that complicated the extrapolation of dose response results from Taiwan to the United States had to do with adjustments for cancer rates among the unexposed population. As described above, available data included person-years at risk and cancer mortality data, stratified by age group, for 42 villages in southwestern Taiwan that had arsenic contamination levels ranging from 10 to over 900 ppb. Clearly, it is possible to fit a dose response model to these data and to use that fitted model to estimate the dose corresponding to a specified excess risk over background. However, since none of the villages was assessed as having a zero level of arsenic contamination, such calculations effectively involve an extrapolation of the fitted model outside the range of observed data. As discussed by Morales et al. (2000), many epidemiologists would argue that the analysis should include appropriate unexposed controls. A relatively common approach would use an analysis based on standardized mortality ratios (SMRs), which involves computing the ratio of observed to expected number of cancer deaths in various exposure-group categories, and modeling these ratios as a function of exposure level. It is straightforward to show (see Breslow and Day, 1987) that the SMR-based approach is asymptotically equivalent to performing a Poisson analysis, as described above, with the population-based data considered as additional data corresponding to an exposure level of 0. In addition to analyses that model data from only the 42 villages, NRC considered analyses that included population data from either the whole of Taiwan, or at least the southwestern region (see Morales et al., 2000, Table 2). Figure 2 shows the estimated village-specific lifetime risks of dying of lung cancer (circles), along with the population-based rate (triangle). The village-specific risks are estimated by fitting a Poisson model like (2), but with the linear exposure term replaced with $h$ village ID specified as a factor variable. The figure illustrates that risks in even the lowest exposed of the 42 study villages were substantially higher than population-based rates. Hence, the fitted dose response curve was sensitive to whether or not the comparison population was included. While dose response models fitted only within the 42 villages yielded relatively consistent results in terms of BMDs, estimated BMDs based on the dataset expanded to include population-based cancer mortality data were much lower, and quite variable. In particular, choice of dose transformation had a relatively strong impact on estimated benchmark dose (see Morales et al., 2000, for further details).

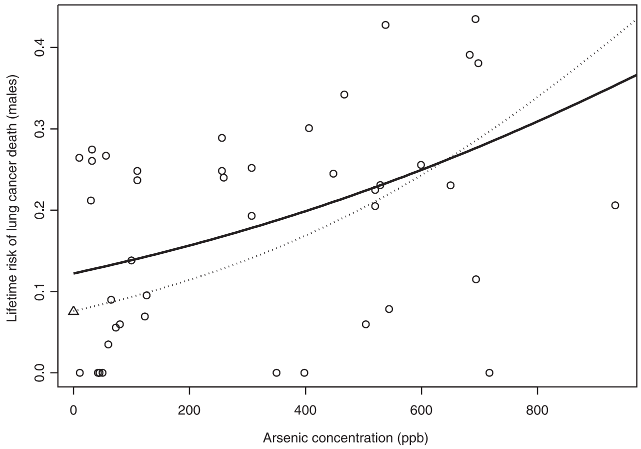

FIG. 2. *Circles show village-specific estimates of lifetime lung cancer death risk* (*males*); *the triangle shows the lifetime risk for the Taiwanese population as a whole. Solid line shows predicted lifetime risk curve when baseline data are used in the analysis*; *dotted line shows predicted curve based only on data from the* 42 *villages*.

*Measurement error.* We have already discussed the potential for bias and increased uncertainty associated with the ecological design of the southwest Taiwanese study. In fact, the situation was even more complicated than we have described in that many of the villages had multiple wells. Figure 3 is a plot of the arsenic levels measured in each of the 42 village wells, ordered by the median level assigned to each village. The figure illustrates that well measurements could be highly variable within a village, with levels varying between 50 and 1752 ppb in one case. Consequently, the levels assigned to each village are likely to be subject to considerable measurement error themselves. Although at first glance it appears that a Berkson measurement error paradigm (see Carroll, Ruppert and Stefanski, 1995) might be appropriate for the analysis of the Taiwanese cancer data, the actual measurement error structure will be hierarchical. While some limited work

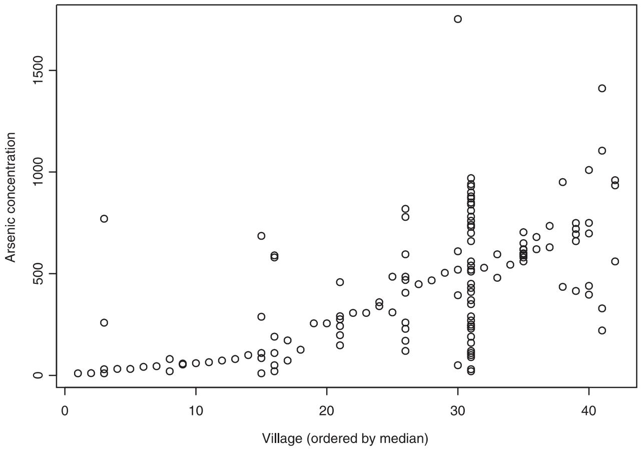

FIG. 3. *Well arsenic levels* (*ppb*), *by village. Villages are ordered according to median arsenic levels. Raw data are available in the Appendix to NRC* (1999) *and also at Statlib.*

has been done on this topic (see Goldstein, 1995, Chapter 10, for some discussion), there is a definite need for further work on the subject of measurement error associated with ecological design and its impact in environmental risk assessment.

*Use of safety factors.* The arsenic risk assessment raises challenging philosophical issues in relation to the use of safety factors. As discussed in Section 2, regulatory levels are usually obtained by dividing estimated benchmark doses by a safety factor of anywhere between 10 and 1,000 to reflect various sources of uncertainty (NRC, 1994). While it can be argued that risk assessments based on epidemiological data do not need to consider uncertainty due to species-to-species variability, other sources of uncertainty (e.g., considerations of sensitive subpopulations) still exist. For arsenic, however, there was a conundrum: assuming a 1% excess risk associated with exposure to 50 ppb of arsenic would suggest that the appropriate MCL be many orders of magnitude lower. While an MCL of 5 ppb is still associated with a high level of risk (approximately 1/1000) by usual EPA standards, this level is close to the limit of detection for arsenic in drinking water. Furthermore, enforcing standards at this level would entail remediation costs in the billions. EPA's final decision to set a new standard of 10 ppb involved a trade-off between health risks and cost. EPA's cost–benefit analysis required them to "monetize" the benefit from the bladder and lung cancers cases avoided. Two different values were used. Cancer cases resulting in death were attributed the "value of statistical life" (VSL), which in 1999 dollars was assumed by EPA to be $6.1 million. A "willingness-to-pay value" (WTP) was used to monetize nonfatal cancer cases. The assumed WTP value was $607,000 in 1999. Despite the highly quantitative nature of cost–benefit analysis, statisticians have not traditionally been active in this area, though it is certainly one where useful contributions could be made. Further interesting reading and references to EPA's arsenic risk assessment can be found at http://www.epa.gov/safewater/arsenic.html.

## 4. GENERAL DISCUSSION

The arsenic case study raises a number of difficult and challenging issues that commonly arise in epidemiologically based risk assessment, including choice of control group, model choice and exposure measurement error. The EPA's risk asessment for arsenic in drinking water ran up against the serious inadequacies of the traditional risk assessment paradigms involving safety factors, as well as some of the difficult challenges of balancing health concerns with economic considerations.

There are no easy answers. In this section, we discuss several areas where some good statistical thinking has the potential for important impact in terms of finding some solutions to these problems.

### 4.1 Biologically Informed Dose Response Curves

Talk to any environmental health scientist and they will tell you that the key to improved risk assessment is identifying appropriate biologically based dose response models. In their 1999 proposed guidelines for cancer risk assessment, the United States Environmental Protection Agency suggests the use of biologically based carcinogenesis models for low-dose risk estimation whenever sufficient evidence is available to support the choice. While in principal this makes perfect sense, it often proves difficult, if not impossible, to implement in practice. When the NAS formed its first committee on arsenic in drinking water, it was hoped that arsenic might be a candidate for using a biologically based model. As the committee deliberated, however, it became clear that there was too much uncertainty and disagreement regarding mechanisms and hence that the default approach, based on a benchmark dose analysis, should be used.

In practice, the term "biologically-based model" has been interpreted in a number of different ways. The general idea is that more accurate dose response models should be obtained by taking account of the various steps going from the original exposure to the final health outcome [see Figure 4, adapted from Schulte (1993)]. How can these ideas be implemented in practice? One approach is to use biological principles to derive a specific dose response model. The multistage model of Armitage and Doll (1954) is a classic example of this approach. Many authors have proposed generalizations of this model. For example, Moolgavkar and Venzon (1979) proposed a two-stage clonal expansion model that assumes that cancer arises from a multistage process that involves proliferation at one of the stages. A problem with such approaches, however, is that they are not generally well identified, except

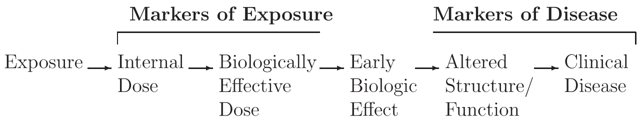

FIG. 4. *Conceptual framework for biologically-based models.*

in relatively simple cases, with only the usual observable data, namely exposure and outcome. To be useful in practice, biologically-based models need to draw on additional data sources. Physiologically based pharmakokinetic (PBPK) models are one such approach, and these have become increasingly popular over the past decade or so (see Smith, 2002).

In contrast to mechanism-based models of carcinogenicity, PBPK models seek improved dose response modelling through the use of more accurate measures of internal dose. PBPK models typically assume that the body comprises several compartments linked together by the circulatory system. A popular choice is a three-compartment model that groups metabolically active organs such as the liver into a well-perfused compartment, which is subject to a relatively high volume of blood flow. Other organs and body structures (including skin, bones, etc.) are grouped into a moderately well-perfused compartment, while body fat forms a poorly perfused compartment. PBPK models can incorporate some mechanistic ideas as well. For example, Smith et al. (2001) use a PBPK model to estimate the rate at which butadiene is metabolized into expoxy butene, which is thought to be one of the primary carcinogenic pathways. A PBPK model is usually characterized through a series of differential equations, each describing the rate of change of concentration of the compound under study in various parts of the body. A particularly appealing feature of the PBPK approach is that it allows the investigator to incorporate information collected from a variety of different studies regarding, for example, flow rates between the various organs. By contributing understanding regarding how the substance travels through the body and by making appropriate assumptions regarding the mechanisms by which an adverse effect occurs, a PBPK model can be invaluable for extrapolating information from one species to another. Because fitting a PBPK model typically requires incorporating information from several different experiments, a Bayesian formulation works particularly well. Wakefield, Smith, Racine-Poon and Gelfand (1994) were among the first to describe the use of a Bayesian hierarchical model to fit a PBPK model, though they considered the relatively simple case where the differential equations characterizing the model could be solved in closed form. Gelman, Bois and Jiang (1996) extended the approach to apply in more complex settings.

Despite their potential, PBPK models have been slow to enter into the mainstream of quantitative risk assessment. In addition to the complexity of fitting them, PBPK models are limited in that, while they are helpful in terms of characterizing internal dose, it is unclear how they can be used to better quantify the link between exposure and the health outcome of interest. There is an important need for the development of new methods that incorporate biological information in a more flexible way. In recent years, environmental scientists have been making rapid progress in identifying relevant biomarkers for a variety of exposure and disease settings [see other papers in the special issue of *Environmental Health Perspectives* introduced by Holian (1996)]. However, statistical methods to incorporate biomarkers into the risk assessment process are lagging behind. While much progress has been made on statistical methods to use biomarkers in the clinical trials context (see, e.g., Prentice, 1989) the focus there tends to be somewhat different, for example, evaluating the usefulness of a biomarker as a surrogate outcome variable. Even so, some of the thinking that has been developed in the clinical trial context could potentially be useful for environmental risk assessment as well. For example, characteristics needed for a biomarker to prove useful for the purpose of dose response modeling might include the following:

- the biomarker is a strong predictor of the outcome of interest, even after adjusting for exposure and other measurable characteristics;
- exposure is a strong predictor of the biomarker.

Ideally, the biomarker should provide a stronger signal than the health outcome alone. For example, biomarkers measured on a continuous scale (e.g., DNA adducts) are likely to be more informative than binary indicators of the presence or absence of rare diseases such as cancer. Ryan et al. (2004) discuss these ideas in a special issue on biomarkers of *Statistical Methods for Medical Research*.

Unfortunately, finding real-world datasets to test out the utility of biomarker data for dose response modeling is difficult. Although many environmental health researchers are actively studying biomarkers in a variety of different disease settings, most studies are focused fairly narrowly on specific mechanistic questions and do not simultaneously collect the exposure and response information needed to construct a dose response model. Many studies have either biomarker data and exposure, or biomarker data and outcome, but not all three together. An exception is an ongoing case–control study in lung cancer in which, in addition to standard measurements of smoking exposure (pack-years, etc.), measurements have been taken on

DNA adducts in blood and tissue, as well as various genetic polymorphisms (e.g., GSTM1) thought to affect susceptibility to smoking-related lung cancer (Zhou et al., 2003). The scientific goal of this study is not to characterize the dose–response relationship between smoking and lung cancer, but rather to understand the genetic basis of lung cancer. As is typical of many such studies, the biomarkers are measured on only a small subset of study subjects. In fact, of 1,842 subjects available for the analysis, only 9 had measurements on all three biomarkers. The sparseness of the observations meant that this particular study provides only a limited opportunity to explore the use of biomarker models for the purpose of dose response modeling. Even so, Ryan et al. (2004) were able to analyze the data using a likelihood-based approach and obtain some useful findings. In particular, they found that DNA adducts were indeed able to explain a good deal of the association between smoking and lung cancer, suggesting that DNA damage is indeed an important pathway.

The lung cancer example raises interesting design challenges. In most epidemiological studies, it will not be possible to measure expensive and labor intensive biomarkers on all study subjects. However, judicious choices about which subjects to measure have the potential to provide considerable improvement in efficiency. To take a heuristic example, if one believes that a particular kind of DNA adduct is part of the causal pathway between exposure to cigarette smoke and the onset of lung cancer, it will be a waste of resources to measure adducts in a large number of nonsmokers. Depending on the design of the main study, it may also make sense to oversample lung cancer cases for the purpose of measuring a biomarker. White (1982) proposed a two-stage case–control design where one starts with a traditional case–control design. Some expensive-or difficult-to-measure covariates are assessed only on a subset of the study subjects, with selection probabilities that depend on the value of other covariates of interest. A number of authors have discussed analytical approaches as well as extensions to the two-stage case–control design. For example, Breslow and Cain (1988) extended the design to allow second stage sampling probabilities to vary according to whether the subject is a case or a control. Further analytic considerations have been discussed by many authors, including Breslow and Holubkov (1997) and Wacholder and Weinberg (1994). Reilly and Pepe (1995) discuss a nonparametric approach to fitting regression models when some covariates are missing and use their formulation to address optimal design considerations as well. Further development of these ideas for dose response modeling would be useful.

### 4.2 Quantifying Uncertainty

While the ideal is of course to base risk assessment on highly accurate, biologically based dose response models, the reality is inevitably a great deal of uncertainty regarding the true relationship between exposure and outcome. Our arsenic case study involved considerable levels of uncertainty from several different sources, but especially in terms of exposure assessment. While it is well known that measurement error can lead to biased estimation of the dose–response relationship and underestimation of the associated uncertainty (see Carroll, Ruppert and Stefanski, 1995), the literature has seen relatively little discussion about the effects of measurement error on benchmark dose calculations. While issues of bias are certainly important, developing better methods to quantify the uncertainty associated with dose response modeling could have significant impact on the field. In the context of the arsenic study, for example, it could be argued that using village-level exposure measures to represent individual exposures should not result in too much bias, since this is an example of so-called Berkson measurement error, which does not lead to bias, at least in the linear model setting. However, the extra variability induced by the uncertainty could be considerable. As a case in point, the NRC arsenic report (NRC, 2001) described an analysis that attempted to account for person-to-person variations in daily drinking water intake. NRC reanalyzed the male lung cancer data using a multiplicative model (2), with $g(x)$ replaced by

$$g(w) = \exp(\beta w),$$

where $w$ represented individual arsenic intake (measured in micrograms per kilogram body weight per day), rather than arsenic concentration in the drinking water. Although $w$ was not observed in the study, a distribution for $w$ could be estimated for each observed village-level arsenic concentration, using an EPA survey that had characterized the population distribution of daily drinking water volumes. The EPA survey suggested that daily drinking rates followed an approximate gamma distribution, with mean 21 milliliters per kilogram body weight per day, and a standard deviation of 15. Keeping in mind that a milliliter is a one one-thousandth of a liter and that parts per billion in water corresponds to micrograms per liter, it follows, for example, that a 50 kg person drinking water from a well contaminated with 100 ppb of arsenic would have a mean daily exposure level of 2.1 ($100 \times 21/1{,}000$) micrograms of arsenic per kilogram body weight per day, with associated standard deviation of 1.5. A Bayesian approach can easily accommodate such data to fit the dose response model in terms of $w$, even though this variable is not directly observable. The Bayesian approach can also be used to generate a posterior distribution of lifetime risk levels associated with each observed concentration of arsenic in drinking water. NRC reported on such an analysis (see Table 5.5 in their report), finding that, as expected, incorporating uncertainty about drinking water rates increased the width of confidence limits substantially. Interestingly, the NRC analysis also suggests that ignoring drinking rate variability also induces some bias in estimated BMDs, despite the fact that the measurement error is of the Berkson class. The explanation for this likely includes the nonlinear nature of the dose–response relationship, as well as the nonnormal distribution of the measurement errors. While a statistician would think of the analysis described here as a classic Bayesian hierarchical analysis, risk assessors might describe it as an example of *probabilistic risk assessment* or perhaps *uncertainty analysis*. A relatively recent development in the risk assessment world, probabilistic risk assessment uses tools such as Monte Carlo simulation to explore the impact of population heterogeneity with respect to exposure, susceptability and basic physiology, to quantify uncertainty in estimation of quantities such as BMDs and lowest-observed-effect levels (LOELs) (see NRC, 1994, Chapter 9). In general, the approach is fairly ad hoc and does not take into account the effect of uncertainty on parameter estimation. With a few exceptions (see Rai, Bartlett, Krewski and Paterson, 2002), few statisticians have worked in the area, which is ripe for further development, especially using a Bayesian formulation. Of course, exposure measurement error is just one of many sources of uncertainty in dose–response modeling. In the case of arsenic, uncertainty regarding the dose–response relationship was itself another major source (Morales et al., 2000). While it is almost certainly true that more accurate exposure measurement would reduce model uncertainty, the biological complexity of most risk assessment settings will lead to uncertainty regarding the true dose–response relationship, even in the context of perfect exposure measurements. Morales (2001) used Bayesian model averaging to incorporate this model uncertainty into risk estimation. Morales' reanalysis of the NRC arsenic data suggests that the Bayesian model averaging approach will lead to similar point estimates as a more traditional approach of basing estimates on the best fitting models. However, confidence limits based on model averaging will be much wider, thus reflecting more appropriately the true uncertainty involved.

## 5. CONCLUSIONS

This paper has used a high-profile case study to highlight some of the interesting and challenging problems that arise in modern environmental health risk assessment, especially when epidemiological data are involved. While the use of epidemiological data avoids many serious criticisms that have been targeted toward the use of animal data for assessing human health risks, other perhaps equally challenging issues arise. This paper has argued that statistical science has the potential to play a significant and central role in guiding the field toward new and appropriate paradigms for environmental health risk assessment. These new paradigms must include an emphasis on not only providing estimates of central tendency, but also quantifying the true population heterogeneity in risk.

An issue of central importance is finding ways not only to improve dose response modeling, but also to accurately characterize the uncertainty in using such models to estimate benchmark doses and other quantities important to regulators. We have argued that Bayesian models provide an ideal framework for this. In our arsenic case study, for example, we saw how a Bayesian hierarchical model could be used to assess the impact of individual variation in an ecological study where exposure levels were measured only at the group (village) level. A particularly appealing aspect of the Bayesian approach in this context is that it facilitates the incorporation of information from a variety of sources, not just the study at hand. In the case of the arsenic analysis, for example, data from an EPA survey was used to characterize the distribution of male and female drinking water rates. Such analyses provide an ideal means of addressing issues of uncertainty, yet at the same time incorporating expert knowledge and other related scientific evidence into the modeling process. Also in the context of our arsenic case study, we described the use of Bayesian model averaging techniques to quantify uncertainty assocated with model choice. There are a number of examples in the literature (see Dominici, Samet and Zeger, 2000) where Bayesian models have been successfully used to synthesize data from several different environmental studies.

While characterizing uncertainty is important, the ideal, of course, is to identify the information and data sources needed to fit accurate dose response models. Developing improved ways to incorporate biological information into dose response modeling is an important area where statisticians have the potential for significant contributions. While significant progress has already been made in the area of pharmacokinetic modeling, such models do not provide all the answers when it comes to risk assessment. Of more value are likely to be hybrid models that combine statistical modeling with biomarker data to find more accurate dose response models relating an exposure of interest to a health outcome. In practice, such models require the incorporation of data from a variety of sources, so that a Bayesian approach will once again provide a natural framework for analysis. Statistical design issues are very important, especially when it comes to incorporating biomarkers. In practice, cost and other practical considerations will make it possible to measure biomarkers on only a subset of study subjects. Providing guidelines for optimal selection of subjects for biomarker assessment is a very important topic worthy of further study.

There are many important statistical problems motivated by environmental health research that this article has not discussed. For example, though we have emphasized the challenges of epidemiologically based risk assessment, many fascinating problems still arise from toxicological studies in animals. For example, the use of genetically altered mice has become particularly popular in toxicology, leading to a need for new statistical methods adapted to this context (see Dunson, 2000).

Although we touched on the topic in our discussion on biomarkers, genetic susceptability to environmentally mediated disease is an important topic that could have been the focus of an entire paper, just on its own. The newly established National Center for Toxicogenomics, established under the auspices of the National Institute of Environmental Health Sciences (see http://www.niehs.nih.gov/nct/home.htm), has as its mission helping environmental scientists to apply modern tools of genomics and proteomics to explore and understand the genetic and metabolic pathways of disease, along with the interactive effect of environmental factors. While great progress has been made in the area of statistical genetics, statistical methods for addressing the problem of detecting gene–environment interactions (Niu, 2002) are still relatively new.

A theme of this paper is that statisticians working in the area of environmental health research need to challenge themselves constantly to think about the big picture and to identify important questions. As in many areas of application, it is easy to become focused on small, technically challenging problems that really do not contribute significantly to the real world. Like many other fields, environmental health is changing rapidly in response to new developments in genomics and other areas of basic science. These changes provide interested statisticians with a great opportunity to make valuable contributions.

## REFERENCES

ABELSON, P. H. (1995). Flaws in risk assessments. *Science* **270** 215.

ALLEN, B. C., KAVLOCK, R. J., KIMMEL, C. A. and FAUSTMAN, E. M. (1994). Dose–response assessment for developmental toxicity. III. Statistical models. *Fundamental and Applied Toxicology* **23** 496–509.

ARMITAGE, P. and DOLL, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *British J. Cancer* **8** 1–12.

BRESLOW, N. E. and CAIN, K. C. (1988). Logistic regression for two-stage case-controlled data. *Biometrika* **75** 11–20.

BRESLOW, N. E. and DAY, N. E. (1987). *Statistical Methods in Cancer Research* **2**. *The Design and Analysis of Cohort Studies*. Oxford Univ. Press.

BRESLOW, N. E. and HOLUBKOV, R. (1997). Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine* **16** 103–116.

BRUMBACK, B. A., COOK, R. J. and RYAN, L. M. (2000). A meta-analysis of case-control and cohort studies with interval-censored exposure data: Application to chorionic villus sampling. *Biostatistics* **1** 203–217.

CARROLL, R. J., RUPPERT, D. and STEFANSKI, L. A. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall, New York.

CHEN, C. J., CHUANG, Y. C., LIN, T. M. and WU, H. Y. (1985). Malignant neoplasms among residents of a blackfoot disease-endemic area in Taiwan: High-arsenic artesian well water and cancers. *Cancer Res.* **45** 5895–5899.

CRUMP, K. S. (1984). A new method for determining allowable daily intakes. *Fundamental and Applied Toxicology* **4** 854–871.

DOMINICI, F., SAMET, J. M. and ZEGER, S. L. (2000). Combining evidence on air pollution and daily mortality from the 20 largest US cities: A hierarchical modelling strategy (with discussion). *J. Roy. Statist. Soc. Ser. A* **163** 263–302.

DUNSON, D. B. (2000). Models for papilloma multiplicity and regression: Applications to transgenic mouse studies. *Appl. Statist.* **49** 19–30.

DUNSON, D. B. and DINSE, G. E. (2002). Bayesian models for multivariate current status data with informative censoring. *Biometrics* **58** 79–88.

FERRECCIO, C., GONZALEZ, C., MILOSAVJLEVIC, V., MARSHALL, G., SANCHA, A. M. and SMITH, A. H. (2000). Lung cancer and arsenic concentrations in drinking water in Chile. *Epidemiology* **11** 673–679.

FINKELSTEIN, D. M. (1991). Modeling the effect of dose on the lifetime tumor rate from an animal carcinogenicity experiment. *Biometrics* **47** 669–680.

FREEDMAN, D. A. and ZEISEL, H. (1988). From mouse to man: The quantitative assessment of cancer risks (with discussion). *Statist. Sci.* **3** 3–56.

GART, J. J., KREWSKI, D., LEE, P. N., TARONE, R. E. and WAHRENDORF, J. (1986). *Statistical Methods in Cancer Research* **3**. *The Design and Analysis of Long-Term Animal Experiments*. Oxford Univ. Press.

GAYLOR, D. W. (1983). The use of safety factors for controlling risk. *J. Toxicology Environmental Health* **11** 329–336.

GELMAN, A., BOIS, F. and JIANG, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions *J. Amer. Statist. Assoc.* **91** 1400–1412.

GOLDSTEIN, H. (1995). *Multilevel Statistical Models*. Arnold, London.

GREENLAND, S. and ROBINS, J. (1994). Ecologic studies—biases, misconceptions, and counterexamples (with discussion). *Amer. J. Epidemiology* **139** 747–771.

HOEL, D. G. and WALBURG, H. E. (1972). Statistical analysis of survival experiments. *J. Nat. Cancer Inst.* **49** 361–372.

HOLIAN, A. (1996). Air toxics: Biomarkers in environmental applications—Overview and summary of recommendations. *Environmental Health Perspectives* **104**(Suppl. 5) 851–855.

KAPLAN, N., HOEL, D., PORTIER, C. and HOGAN, M. (1987). An evaluation of the safety factor approach in risk assessment. In *Banbury Report 26*: *Developmental Toxicology*: *Mechanisms and Risk* (J. A. McLachlan, R. M. Pratt and C. L. Markert, eds.) 335–346. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

KIMMEL, C. A. and GAYLOR, D. W. (1988). Issues in qualitative and quantitative risk analysis for developmental toxicology. *Risk Anal.* **8**(1) 15–20.

KODELL, R. and NELSON, C. (1980). An illness–death model for the study of the carcinogenic process using survival/sacrifice data. *Biometrics* **36** 267–277.

LAIRD, N. M. and OLIVIER, D. C. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *J. Amer. Statist. Assoc.* **76** 231–240.

LEISENRING, W. and RYAN, L. M. (1992). Statistical properties of the NOAEL. *Regulatory Toxicology and Pharmacology* **15** 161–171.

LIN, T., GOLD, L. S. and FREEDMAN, D. (1995). Carcinogenicity tests and interspecies concordance. *Statist. Sci.* **10** 337–353.

MANCUSO, J. Y., AHN, H., CHEN, J. J. and MANCUSO, J. P. (2002). Age-adjusted exact trend tests in the event of rare occurrences. *Biometrics* **58** 403–412.

MOELLER, D. (1992). *Environmental Health*. Harvard Univ. Press. Cambridge, MA.

MOOLGAVKAR, S. H. and VENZON, D. J. (1979). Two-event models for carcinogenesis: Incidence curves for childhood and adult tumors. *Math. Biosci.* **47** 55–77.

MORALES, K. H. (2001). Statistical issues in quantitative risk assessment with application to arsenic in drinking water. Ph.D. dissertation, Harvard School of Public Health.

MORALES, K. H., RYAN, L. M., KUO, T.-L., WU, M.-M. and CHEN, C.-J. (2000). Risk of internal cancers from arsenic in drinking water. *Environmental Health Perspectives* **108** 655–661.

MORGAN, B. J. T. (1992). *Analysis of Quantal Response Data*. Chapman and Hall, London.

NATIONAL RESEARCH COUNCIL (NRC) (1983). *Risk Assessment in the Federal Government*: *Managing the Process*. National Academy Press, Washington, DC.

NATIONAL RESEARCH COUNCIL (NRC) (1988). *Health Risks of Radon and Other Internally Deposited Alpha-Emitters: BEIR IV. Committee on the Biological Effects of Ionizing Radiations*. National Academy Press, Washington, DC.

NATIONAL RESEARCH COUNCIL (NRC) (1991). *Environmental Epidemiology*. National Academy Press, Washington, DC.

NATIONAL RESEARCH COUNCIL (NRC) (1994). *Science and Judgement in Risk Assessment*. National Academy Press, Washington, DC.

NATIONAL RESEARCH COUNCIL (NRC) (1999). *Arsenic in Drinking Water*. National Academy Press, Washington DC.

NATIONAL RESEARCH COUNCIL (NRC) (2001). *Arsenic in Drinking Water*: *2001 Update*. National Academy Press, Washington, DC.

NIU, T. (2002). Gene–environment interaction. In *Encyclopedia of Environmetrics* **2** 848–851. Wiley, New York.

OLDEN, K. and GUTHRIE, J. (1996). Air toxics regulatory issues facing urban settings. *Environmental Health Perspectives* **104**(Suppl. 5) 857–860.

PIEGORSCH, W. and BAILER, A. J. (1997). *Statistics for Environmental Biology and Toxicology*. CRC Press, Boca Raton, FL.

PRENTICE, R. L. (1989). Surrogate endpoints in clinical trials—Definitions and operational criteria. *Statistics in Medicine* **8** 431–440.

PRENTICE, R. L. and SHEPPARD, L. (1995). Aggregate data studies of disease risk factors. *Biometrika* **82** 113–125.

RAI, S. N., BARTLETT, S., KREWSKI, D. and PATERSON, J. (2002). The use of probabilistic risk assessment in establishing drinking water quality objectives. *Human and Ecological Risk Assessment* **8** 493–509.

REILLY, M. and PEPE, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82** 299–314.

RYAN, L. M., HUANG, W., THURSTON, S. W., KELSEY, K. T., WIENCKE, J. K. and CHRISTIANI, D. C. (2004). On the use of biomarkers for environmental health research. *Stat. Methods Med. Res.* To appear.

SCHULTE, P. A. (1993). Use of biological markers in occupational health research and practice. *J. Toxicology and Environmental Health* **40** 359–366.

SMITH, T. J. (2002). Issues in exposure and dose assessment for epidemiology and risk assessment. *Human and Ecological Risk Assessment* **8** 1267–1293.

SMITH, T. J., LIN, Y. S., MEZZETTI, M., BOIS, F. Y., KELSEY, K. and IBRAHIM, J. (2001). Genetic and dietary factors affecting human metabolism of 1,3-butadiene. *Chemico–Biological Interactions* **135** 407–428.

TSENG, W. P., CHU, H. M., HOW, S. W., FONG, J. M., LIN, C. S. and YEH, S. (1968). Prevalence of skin cancer in an endemic area of chronic arsenicism in Taiwan. *J. Nat. Cancer Inst.* **40** 453–463.

U.S. ENVIRONMENTAL PROTECTION AGENCY (U.S. EPA). (1988). Special report on inorganic arsenic: Skin cancer; nutritional essentiality. EPA 625/3-87/013, U.S. Environmental Protection Agency, Washington, DC.

U.S. ENVIRONMENTAL PROTECTION AGENCY (U.S. EPA). (1999). Guidelines for carcinogen risk assessment. Review Draft, NCEA-F-0644, July 1999, Risk Assessment Forum. Available at http://www.epa.gov/ncea/raf/cancer.htm.

U.S. ENVIRONMENTAL PROTECTION AGENCY (U.S. EPA). (2002). Health assessment document for 1,3-butadiene. EPA/600/P-98/001, Office of Research and Development, Washington, DC.

WACHOLDER, S. and WEINBERG, C. R. (1994). Flexible maximum-likelihood methods for assessing joint effects in case–control studies with complex sampling. *Biometrics* **50** 350–357.

WAKEFIELD, J. (2003). Sensitivity analyses for ecological regression. *Biometrics* **59** 9–17.

WAKEFIELD, J. C., SMITH, A. F. M., RACINE-POON, A. and GELFAND, A. E. (1994). Bayesian analysis of linear and nonlinear population models by using the Gibbs sampler. *Appl. Statist.* **43** 201–221.

WHITE, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *Amer. J. Epidemiology* **115** 119–128.

WRIGHT, C., LOPIPERO, P. and SMITH, A. (1997). Meta analysis and risk assessment. In *Topics in Environmental Epidemiology* (K. Steenland and D. Savitz, eds.) 28–63. Oxford Univ. Press.

ZHOU, W., LIU, G., MILLER, D. P., THURSTON, S. W., XU, L. L., WAIN, J. C., LYNCH, T. J., SU, L. and CHRISTIANI, D. C. (2003). Polymorphisms in the DNA repair genes XRCC1 and ERCC2, smoking, and lung cancer risk. *Cancer Epidemiology*, *Biomarkers and Prevention* **12** 359–365.