Comment

BRIAN MACMAHON*

1. INTRODUCTION

The article by Lagakos, Wessen, and Zelen is the third public report of a body of work that has become known as "the Woburn Study" (Lagakos, Wessen, and Zelen 1984a,b). For purposes of discussion we must understand that this is not a single study. It is composed of two distinct studies—the leukemia study and the health survey (the latter of which must itself, for purposes of inference, be considered in two parts that deal with reproductive events and are concerned with childhood illness, respectively). These studies and their components are linked only by a complex, though still imprecise, estimate of exposure to the water from two contiguous Woburn wells (G and H). This common feature should not be allowed to blur the distinctions between the studies, because the strengths and weaknesses of each do not necessarily carry over into the others, although they tend to do so in the public mind and legal arena. For example, the fact that a pregnancy occupies a specific place in time allows advantage to be taken of the fact that wells G and H did not contribute to the Woburn water supply in equal proportion throughout the study period. This is a strength that applies to analyses of exposure to that water in relation to outcomes of pregnancy but that does not carry over into analyses relating to the leukemia cluster or to childhood illnesses. I will address in turn what seem to me three distinct issues.

2. CHILDHOOD LEUKEMIA

Before the study of Lagakos et al. it was known that there was an excess of childhood leukemia in Woburn between 1969 and 1979 and that the cluster was centered in East Woburn. Their study has confirmed the existence of an excess rate of childhood leukemia in East Woburn and supplied the new observation that the excess risk for Woburn continued beyond 1979, although, curiously, in West rather than East Woburn. The central issue is whether the investigators have gone beyond this geographic clustering and provided evidence, as they suggest, that it is the water supplied to East Woburn by wells G and H that is responsible for at least part of the excess. I believe that they have not. My reasons include the following:

1. In relating leukemia risk to estimated exposures to the water of wells G and H, two measures ("metrics") of exposure are used, one dichotomous (some/none) and the other continuous (cumulative exposure). To be in the "some" exposure category, a child has only to reside in an area served by the wells in some year when they were pumping. The wells were in operation for most of the study period, although the proportion of G and H water in the supply of a particular district varied over time. On the other hand, the cumulative exposure takes account of changes in residence over the child's lifetime and the proportion of G and H water in the supply to the district of his home in each year of life; it appears, therefore, to be a more accurate measure of exposure than the dichotomy none/some. Indeed, if it were not, why bother with it? If the exposure to G and H water were causally related to risk of leukemia, one would expect that the more accurate measure of it would yield the stronger measure of relationship. In fact, the contrary is the case. The p values given by the authors are lower for the some/none dichotomy (.02) than for the cumulative exposure analysis (.03). This can only mean that the data are inconsistent with an underlying linear (on the log scale) relationship between cumulative exposure and rate of disease. Risk ratios (RR) also suggest that the underlying relationship is markedly nonlinear. Thus the RR for some exposure compared to none is 3.03 (antilog 1.11). The corresponding RR computed from the cumulative exposure analysis can be computed from the authors' Table 2, which gives, for each risk set sample, the expected cumulative exposure and the proportion of the risk set exposed. Since all of the cumulative exposure occurs, by definition, among the exposed, the mean cumulative exposure *among the exposed* over all risk sets can be computed as 2.12. The antilog of 2.12 × .33 (the regression coefficient), or 2.02, is an estimate of the relative risk associated with the average exposure experienced by exposed individuals. It is actually lower than the 3.03 estimated from the simple some/none dichotomy. To obtain the same estimate of RR from the cumulative exposure analysis as was derived from the some/none dichotomy the mean exposure among exposed subjects would have to be 3.36; only two of the nine exposed cases have values as high as or higher than this. The authors note that ". . . there are two few cases to be confident of which, if either (of these two methods of analysis), best describes this relationship" (p. 588). It may be true that it cannot be determined whether the difference between the results using the two methods is due to chance, but it is also true that, relative to use of the primarily geographically determined crude measure, use of the complex measure of exposure does not strengthen the association either in terms of risk estimation or statistical testing. The complex measure is, therefore, at best uninformative.

2. Another cruder measure of exposure than cumulative exposure is referred to in the authors' final paragraph preceding the Discussion. Woburn is partitioned into two geographic zones according to the area of coverage of wells G and H under average pumping conditions. We are told that this analysis resulted in "the same significant associations" as the original analyses. We are given no numerical

* Brian MacMahon is Henry Pickering Walcott Professor, Department of Epidemiology, School of Public Health, Harvard University, Boston, MA 02115.

Journal of the American Statistical Association
September 1986, Vol. 81, No. 395, Applications

estimates but led to infer that this degradation of the measurement of exposure to G and H water was also without notable effects on the significance of the association.

3. None of the known contaminants of G and H water is a known leukemogen, in man or laboratory animals, and although their presence in drinking water is considered undesirable, since several are animal carcinogens at high exposure levels, they are in such low concentrations in the well water that it would involve a major revision of our ideas about chemical carcinogenesis to believe that they are indeed causally associated with a doubling of leukemia risk.

4. The association with G and H water, even if causal, explains only about half the excess of childhood leukemia in Woburn. The necessity to invoke other causes for the remainder makes less attractive the idea that half the excess is caused by G and H water.

In their presentation of the leukemia data the authors are thorough and open. Thus one can see from their Table 2 that among the seven new cases after 1979 only one had any exposure to G and H water and that that case had an exposure almost as low as the lowest exposure of any previous exposed case, and in their text they point to the shift in the cluster after 1979 to West Woburn, to which G and H water has never been pumped. But this and similar awkwardnesses do not seem to be adequately weighed in their bottom-line emphasis on the "association" between access to water from the two wells and risk of childhood leukemia. A more balanced presentation of the many possible interpretations of the leukemia cluster in Woburn, including the small though still real possibility that it might be due to chance, appears in the *Final Report* of the Woburn Advisory Panel to the Massachusetts Department of Public Health (1985).

3. ADVERSE OUTCOMES OF PREGNANCY

Data on adverse outcomes of pregnancy and childhood illnesses come from the so-called health survey, the imperfections of which (low response rate, strong opportunity for response bias, poor information quality) can only be evaluated partially by the means employed by the investigators. For example, six of the volunteer interviewers were from the families of the leukemia cases and some of them were involved in litigation concerning the contaminated water. One wonders why such volunteers were accepted as interviewers in the first place. The results obtained by these six, however, were examined and found not to differ from those of others and "virtually identical study results would have been obtained if the data from these six interviewers were omitted" (p. 593). The latter is hardly surprising, since these were only 6 out of a total of 235 active interviewers. This evaluation, however, takes no account of the fact that the members of FACE also participated in the training of interviewers (Lagakos et al. 1984a) and that their influence extends beyond the six mothers of leukemia cases. Further, even though care was taken to see that the interviewers were blind as to area of residence of the interviewees, the subjects themselves certainly were not and the opportunity for response bias stemming from the respondents rather than the interviewers seems considerable.

With respect to adverse outcomes of pregnancy, as already noted, the facts that the time of occurrence of the causes of these outcomes can be approximately identified and that exposure of the population to G and H water varied over time make a measure derived from both place and time intuitively more appealing. Some use of these facts is made in the authors' Tables 7 and 8, although because of nonconcordance of time periods and geographic zones (East and West Woburn or Zones A–C and D–E), I am unable to extract from these tables the numbers that would satisfy me empirically that the complex measure is a better measure than a simple cross-tabulation by year and broad geographic area (e.g., East and West Woburn).

There are two characteristics of this section of the report that are of particular concern: (a) the failure of the investigators to take adequate cognizance of the fact that in a study in which many outcomes are related to a suspected exposure some statistically significant associations will arise as the result of chance and (b) the biologic implausibility of the significant associations that do arise or in some instances are created by meaningless groupings of diagnoses.

With respect to the first point, one must recognize that there were no a priori hypotheses in this component of the study and that the interpretation of "statistical significance" and the meaning of the numerous p values are issues on which there could be a great deal of disagreement. Not being a statistician, I am not about to propose the appropriate solution to the multiple testing problem in this context. Nevertheless, unless the investigators think the problem does not exist here—and I think that they would be wrong to take that position—they should have provided some guidance as to the meaning of the occurrence of 2 or 3 statistically significant associations of G and H water with congenital malformations when several hundred categories were or could have been tested.

Regarding the second point, among adverse pregnancy outcomes significant associations were found with perinatal deaths since 1970 and with certain categories of congenital malformation. For perinatal deaths, the authors note that "the positive association with G and H exposure is primarily due to the three events in the '.51–1.00' exposure interval; these were stillbirths and occurred in 1977–1978 to women with G and H exposure scores of .94, .94, and 1.00" (p. 589). The unadjusted data are given in Table 5. It is remarkable that of the four perinatal deaths in 1970–1982 with any exposure at all three occurred at the very top of the range of exposure and within a time period of 2 years. Indeed for exposures less than .51, there is only one perinatal death in 193 pregnancies, giving a crude rate per thousand of 5.2—slightly less than that for nonexposed pregnancies. It is only the extreme exposure scores accumulated by these three deaths that allows a "significant" relationship to be developed with the cumulative exposure score. Although the causes of stillbirths are not always clearly identifiable, it would have seemed desirable in in-

terpreting this strange finding to see whether any commonality of causal pathology could be found—either in fetal death certificates or in hospital records.

Significant associations were found with two categories of congenital defect—eye/ear anomalies and CNS/chromosomal/oral cleft anomalies. Both of these categories are created by the investigators out of whole cloth. With respect to the first, it is stated that "Medically diagnosed congenital anomalies were grouped according to the involved organ or system using the International Classification of Disease (ICD) codes" (p. 585). Eye/ear defects are said to be one of these groups. But of the 18 cases of defect classified by the investigators to this group 14 are not classified as congenital anomalies by the ICD. These 14 appear in the eye disease section of the ICD, not as congenital anomalies. The authors might take issue with ICD in some instances (e.g., it is curious that ICD regards born deaf as a congenital anomaly but not born blind); the authors might with appropriate explanation have persuaded the reader of the reasonableness of including congenital blindness here. But there is no justification whatsoever for regarding as congenital anomalies such diagnoses as amblyopia, strabismus, "eyes severely crossed"—which account for 7 of the 18 cases in the category. Of the seven cases that should not have been in the category, four were exposed and three were unexposed to G and H water; in those categories of malformation said not to show an association with G and H water the ratio of exposed to unexposed is about 1:7. At the very least, the origin of this category is inadequately explained; at worst it gives the appearance of gerrymandering.

The other category of defects showing a significant association combines CNS with chromosomal defects and oral clefts. The basis for this category is that the investigators "could find (for these defects) *assertions* in the literature of *potential* links with chemicals, pesticides, or trace elements" (p. 585; parentheses and emphasis added). Note that the word is *assertions,* not evidence. I can only comment that the investigators did not look far, because the list of defects that have been *asserted* to be linked to one or other of these compounds is a great deal longer than the one they produced. The grouping of CNS and chromosomal defects with oral clefts does not constitute an a priori grouping that has any foundation in theory or empiric observation. The error is compounded by the fact that 10 of the 27 cases in this category (mental retardation and cerebral palsy) are also not classified as congenital anomalies by the ICD, again without explanation or apology.

The investigators analyze the three diagnoses in the last category individually. They state that the association is present for CNS and chromosomal defects, but not for oral clefts (there were in fact no exposed infants with oral clefts). Only 5 of the 15 infants assigned to CNS defect, however, should be so assigned according to ICD. The other diagnoses (mental retardation and cerebral palsy) are not regarded as congenital anomalies either by ICD or investigators generally. It is not stated whether for these individual malformations the associations are statistically significant. On the basis of the numbers, it would appear not, and even less so if the incorrectly assigned cases are excluded.

4. OTHER CHILDHOOD DISORDERS

Significant associations are reported for diseases of the kidney or urinary tract (said to be primarily "kidney or urinary infections," although of which organ is not stated) and for lung/respiratory tract disorders (said to be "mostly asthma, chronic bronchitis, or pneumonia"). These categories of disorders share with leukemia the difficulty of identifying time of operation of causal factors, and we are back to the issue of whether anything more than a geographical clustering has been shown. The authors' Table 8 addresses this issue to some extent but is heavily weighted by data from the period after 1979 when the wells were closed. It is of course a relevant observation that observed and expected numbers of illnesses were similar after the wells were closed, but it does not answer the question of whether *during the period when these conditions were in excess* (assuming that there was such a period) the cluster was better delineated by a measure of exposure to G and H water than on a simple geographic basis. The data that enable the reader to evaluate this issue for leukemia—and persuaded me that it was not—are not given for the childhood disorders.

5. CONCLUSION

In this series of studies the application of complex mathematics has served to confuse rather than illuminate public health issues principally because of inadequate concern for their biologic aspects. Greater complexity of measurement of exposure has been thought to be necessarily better and that has turned out not to be the case—either that, or the wrong exposure has been measured. The health survey component has severe methodologic weaknesses that can be evaluated only partially. There has been too little attention to the existing epidemiological, experimental, and medical knowledge that should go into creating categories of disease for etiologic study. The investigators have brought forward some associations that may warrant further investigation in other geographic areas, but to state that there was "a consistent and recurring pattern of positive associations with availability of water from wells G and H" as the authors have in a previous publication (Lagakos et al. 1984a, p. 1), is to grossly overinterpret the data. In the present article the authors have moderated their interpretation (in the right direction), but there remains a way to go to where I stand.

ADDITIONAL REFERENCES

Lagakos, S. W., Wessen, B. J., and Zelen, M. (1984a), "The Woburn Health Study: An Analysis of Reproductive and Childhood Disorders and Their Relation to Environmental Contamination," technical report, Boston, MA: Harvard School of Public Health, Dept. of Biostatistics.

——— (1984b), "An Analysis of Contaminated Well Water and Health Effects in Woburn, Massachusetts," SIMS Technical Report 3, SIAM Institute for Mathematics and Society.

Woburn Advisory Panel to the Massachusetts Department of Public Health (1985), *Final Report,* Dept. of Health, Commonwealth of Massachusetts.

Comment

ROSS L. PRENTICE*

I would like to begin by complimenting the authors on their considerable effort in conducting this study, especially since it was carried out, I suspect, on a shoestring budget. Furthermore, the authors provide a helpful discussion of potential biases in the study. My comments are mostly directed to the strength of conclusions that seem merited at the end of this study, in view of potential biases as well as other more purely statistical issues.

First consider childhood leukemia. The authors use tests derived from Cox's regression model to assert a significant association between childhood leukemia incidence in Woburn and indexes of exposure to contaminated wells G and H located in East Woburn. Specifically, two test statistics, one based on a measure of cumulative exposure and one based on an indicator variable for any prior exposure, yielded two-sided significance levels of .06 and .04, respectively (two-sided significance tests will be used throughout this discussion in order to avoid the somewhat anomalous situation of Sec. 4.3, wherein a negative association with exposure to water from G and H wells cannot be declared significant regardless of its strength). How confident should we be that the data merit the conclusion that an association, significant at conventional levels, has been found?

First consider some rather technical aspects. The reported test statistics are based on 17 childhood leukemia occurrences, 9 among children with some prediagnostic exposure to water from wells G and H, and 8 without such exposure. How accurate are asymptotic significance levels in the presence of such small numbers of events? For example, Hoel and Jennrich (1985) considered a similar example in which a log-rank trend test based on 20 events had significance level estimates of .0007 using the usual asymptotic distributional approximations, whereas the actual significance level, as determined by simulation, appeared to be in the vicinity of .01. Might it be plausible that the actual significance levels that attend these tests are in the vicinity of .10, say, rather than .05? A related point concerns the substantial sensitivity of these test statistics to the deletion or addition of single observations. As the most extreme example, if case 12 is omitted from the calculation, the significance level for the cumulative exposure-based test rises from .06 to .42. Even the inclusion of a single omitted unexposed case has noticeable effect on calculated significance levels. For example, the addition of an unexposed case with the same characteristics (year of birth, year of diagnosis, residency period) as case 12 would increase the significance level of the cumulative exposure-based test from .06 to .14 and the exposure indicator-based test from .04 to .07. Incidentally, if exposure to water from G and H is causally related to childhood leukemia, one might have hoped that a judicious choice of exposure index would have led to a more highly significant test than does that based on a simple binary exposure indicator.

The test statistic sensitivity to individual data points lends importance to a rather fundamental concern in respect to leukemia case ascertainment. The authors do not provide evidence that any systematic approach to the identification of childhood leukemia cases was undertaken. They mention a list prepared by a citizens group, a subsequent (1981) report that included 12 cases diagnosed in 1969–1973, and an updated series of 20 cases for the present study. It helps somewhat to know that these 20 are all the cases identified by the state and Dana-Farber Cancer Institute/Children's Hospital registries, but to what extent are these sources intended to be population-based? Is it possible that in an atmosphere of environmental concern and litigation the intensity of timely case ascertainment may have been greater, say, in East Woburn than in West Woburn? Do all of the 20 diagnoses satisfy standardized diagnostic criteria as used, for example, by the SEER system?

The present study apparently derived strong motivation from the preceding (1981) study, which involved 12 childhood leukemia cases versus 5.3 expected. We might be particularly concerned about comparability of case ascertainment within and outside of Woburn, and we may be interested in how an emphasis on childhood leukemia developed (e.g., was a "cluster" first identified?) in interpreting the significance level ($p = .008$) in that particular study. Of more direct consequence for the present analysis is the observation that "the leukemia excess was accounted for primarily by six cases occurring in one of the town's six census tracts" (p. 583). The census tract in question is located on the eastern side of Woburn, so some knowledge of an east versus west difference in leukemia incidence appears to have preceded the present investigation. A somewhat cynical view is then that any factor that distinguishes East Woburn from West Woburn would be expected to exhibit some relationship to childhood leukemia incidence. The authors note, in their discussion section, a number of factors that are similar between East and West Woburn. With the background that preceded this study, however, it seems important to demonstrate that water source, rather than a myriad of other factors that may distinguish East and West Woburn, is specifically associated with leukemia occurrence. Such demonstration can, of course, only be approximated in any observational study, but the fact that the aforementioned score tests accommodate only year of birth as a potential confounding factor seems a rather severe limitation in interpreting the results of this study. One wonders, for example, how much significance levels may be affected by even such a simple modification as stratification on both year of birth and census tract.

* Ross L. Prentice is Associate Director for the Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98104.

Journal of the American Statistical Association
September 1986, Vol. 81, No. 395, Applications

A final detail concerning the childhood leukemia test statistics is that they are technically not based on any partial likelihood function, since the noncases represent only a subset of the cohort from which the cases arise. The risk-set sampling approach referred to by the authors does not cover this application, since risk-set samples must be selected independently and randomly from the entire cohort at risk at each failure time to generate a partial likelihood. The case-cohort sampling approach described in Prentice (1986) does apply provided individuals included in the sample survey are a random sample of the entire cohort. The test statistic variance estimation then, however, includes covariance terms not present for partial likelihood score tests. In this application, however, such terms are almost certainly negligible.

Consider now, more briefly, the results on adverse pregnancy outcomes, childhood disorders, and exposure to water from wells G and H. The authors appear to have worked diligently to avoid bias and to collect the best quality data that could be obtained within the resource constraints of the project. In view of a high level of citizen concern, however, one cannot help but wonder to what extent the telephone interviewers remained blinded as to the interviewees' residence location, or more generally, the extent to which the results would be the same if professional interviewers had been employed. (The authors do comment that the exclusion of interviews conducted by family members of the leukemia cases would not materially affect their results.) Along the same lines one wonders whether the closure of wells G and H and related publicity could have sensitized the "exposed" families toward more complete reporting of adverse pregnancy outcomes or childhood disorders. Unfortunately, available resources did not allow possible differential underreporting of adverse health effects between exposed and unexposed respondents to be examined.

With eight adverse pregnancy end points under consideration one might worry about the extent to which the interpretation of individual significance should be adjusted to acknowledge multiple testing issues. For example, if such tests were independent there would be a probability of .34 of at least one of the eight tests being significant at the .05 level. The observation that two of the eight tests were significant at the .01 level, however, is not wholly explainable in terms of the multiplicity of tests. The six adverse pregnancy outcome tests that were not significant each gave rise to two additional tests of association, one for the years 1960–1969 and one for 1970–1982. That one of these 12 (perinatal death, 1970–1982) should prove "significant" seems readily explainable on the basis of chance alone.

Multiple testing considerations seem particularly pertinent in respect to childhood disorders. Of the nine disorder categories examined, only one gives a (two-sided) significance level less than .05. That test suggests a positive association between kidney/urinary tract disorders and exposure to water from wells G and H ($p = .04$). The probability of one or more significant associations at the .05 level in nine independent tests is .37. Moreover, the next most highly significant association ($p = .06$) suggests a negative relationship between heart and blood pressure disorders and exposure to G and H water. A reasonable summary of these data would seem to be that there is little evidence for an association between childhood disorders and exposure to water from the two wells.

The data on rate changes since the closure of wells G and H are interesting, but it is difficult to know how formally they can be interpreted. It is tempting to interpret the absence of reported adverse pregnancy outcomes during 1980–1982 (Table 7) as supportive of a reduction in risk following closure of the wells; one's willingness to do so, however, should be tempered by the fact that childhood leukemia rates in West Woburn increased to four in the 4 years after well closure from four in 16 years prior to well closure. The principle seems to be that one should not read too much into observations involving very small numbers of events.

In summary, this is an excellent article for discussion. Various aspects of the activity preceding the study, the conduct of the study, and the features of the data place the assessment of the strength of evidence well outside of the rather specialized circumstances for which we know how to calculate, or approximate, significance levels. Based on the information supplied in this manuscript, my somewhat informal summary of strength of evidence is as follows: There is suggestive, but not strong, evidence for an association between exposure to wells G and H and childhood leukemia incidence. In view of a rather limited accommodation of potential confounding factors the issue of causality of such an association appears to merit further consideration. As the authors note in their discussion, literature on the relationship between pollutants found in wells G and H and leukemia, or cancer more generally, may help ascribe a plausible causal role to ingestion of water from these sources. The data on adverse pregnancy outcomes appear to be somewhat stronger, though the fact that the significant associations involved combinations of rather distinct disease entities (eye/ear anomalies; CNS/chromosomal/oral cleft anomalies) is somewhat distracting. There appears to be little reason to suppose that water source is associated with the childhood disorders considered. For each of the three end point categories considered, possible differential case ascertainment is a substantial concern in interpreting the data.

I will be interested in the authors' response to this summary. Undoubtedly, they are aware of many aspects of this study that are unknown to me and that may affect their interpretation of the data. Let me close by thanking the authors for allowing their work to be commented on in this forum.

REFERENCES

Hoel, D. G., and Jennrich, R. I. (1985), "Life Table Analysis With Small Numbers of Cases: An Example-Multiple Myeloma in Hiroshima and Nagasaki," *Journal of Statistical Computation and Simulation*, 20, 311–322.

Prentice, R. L. (1986), "A Case-Cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials," *Biometrika*, 73, 1–11.

the attention of the scientific community to the problem of documenting the adverse health effects of low-level environmental contamination, the authors have done a service. We hope that their initial work will be followed up by well-funded, carefully designed studies of Woburn and other communities exposed to low-level, environmental contamination.

ADDITIONAL REFERENCES

California Department of Health Services (1985), *Pregnancy Outcomes in Santa Clara County, 1980–1982: Reports of Two Epidemiological Studies,* State of California Publications Section, Pub. 7540-958-1301-5.

Hill, Bradford (1971), *Principles of Medical Statistics* (9th ed.), London: Lancet Ltd., p. 245.

Robins, J. M., Cullen, M. R., and Welch, L. S. (in press), "Improved Methods for Discerning Health Impacts of Current Technologies," in *Environmental Impacts on Human Health: An Agenda for Long-Term Research and Development,* ed. Sidney Dragen, New York: Praeger.

Robins, J. M., Landrigan, P. J., Robins, T. G., and Fine, L. (1985), "Decision Making Under Certainty in the Setting of Environmental Health Regulation," *Journal of Public Health Policy,* 6, 322–328.

Stringfellow Health Effects Study: An Epidemiological Health Survey of Residents of Glen Avon and Rubidoux, California (1986), Report prepared by Dean Baker and Sander Greenland for the California Department of Health Services, Los Angeles.

Comment

ALICE S. WHITTEMORE*

The authors of this interesting article have attacked a difficult problem with limited funds, and in a highly politicized climate. Their findings are disturbing. Perhaps most worrisome, children from households served by the contaminated wells had significantly greater leukemia rates than did children from other households. This association will strengthen the lawsuit brought against two major corporations by the families of seven children, victims of leukemia that their parents say was caused by industrial pollution of drinking water.

But is the association causal?

The authors are properly careful to warn us against leaping to that conclusion. Their analysis and discussion, however, sidestep the causality issue and fall short of providing the framework we need to wrestle with it.

The struggle to distinguish causal from noncausal relationships predates the disciplines of epidemiology and biostatistics. In 1840, Jakob Henle published postulates for evaluating a causal relationship between a new infectious agent and a clinical disease [see Henle (1938) for an English translation]. His pupil, Robert Koch, developed these postulates and presented them in 1890 before the International Medical Congress in Berlin (Koch 1890). More recently, Sir Austin Bradford Hill (1965) modified them for epidemiological studies of environmental agents and noninfectious diseases. Applied to the present problem, Hill's postulates require that the association between contaminants in drinking water and the leukemias should have (1) strength, (2) consistency, (3) specificity, (4) temporality, (5) a dose–response relationship, (6) biological plausibility, (7) coherence, (8) experiment, and (9) analogy. It is important to assess the drinking water-leukemia association, according to each of these criteria.

1. The *strength* of the association describes the magnitude of the disease rate in the exposed versus the unexposed. In Section 4.1 we calculate that children from households served by the wells had $\exp(1.11) = 3.03$ times the rate of leukemia incidence than did other Woburn children in the survey. The article provides us with no confidence limits for this estimate, so it is difficult to interpret it as a measure of strength. As it stands, it is moderately strong: weaker than the lung cancer rate ratios of 10 or more experienced by lifelong heavy smokers relative to nonsmokers, and stronger than the heart disease rate ratios of two or so associated with smoking. The authors remind us that if a contaminant in the water did cause some of the leukemias, the large error with which they measured exposure to that contaminant produces a bias toward unity in the observed rate ratio.

2. The article does not tell us much about the *consistency* of the leukemia-drinking water association in relation to other epidemiological studies. It is provocative that unexplained clusters of childhood leukemias have been found in many parts of the world (e.g., Knox 1964; Pinkel and Nefzger 1959). Could it be that unmeasured contaminants in the drinking water caused them? The evidence from other data does not support such a conclusion for the present study. At least two investigations of contaminants in drinking water and site-specific cancers found no association with adult-onset leukemia (Gottlieb, Carr, and Clarkson 1982; Wilkins and Comstock 1981). Thus the current findings are not consistent with the results of other studies.

3. The *specificity* of an association is a measure of its uniqueness. For example, nothing other than exposure to polyvinyl chloride monomer has been associated with angiosarcoma of the liver, and conversely, this compound has not been strongly associated with other diseases. By contrast, exposures to benzene, ionizing radiation, and certain viruses have been associated with one or more of the adult and childhood leukemias, and the authors tell us that other

* Alice S. Whittemore is Professor (Research), Department of Family, Community and Preventive Medicine, Stanford University School of Medicine, Stanford, CA 94305.

factors must be responsible for the Woburn leukemia excess. Conversely, we have seen that drinking water from contaminated Woburn wells has also been associated with perinatal deaths, congenital anomalies, and certain childhood disorders. This lack of specificity mitigates against a causal explanation for the leukemia association, particularly in light of the biological implausibility of some of the other associations.

4. In assessing the *temporality* of the association, we need to know that exposures to the contaminated well water preceded the onset of the leukemias. This is difficult because we do not know when the wells were first contaminated, and we do not know the actual onset time of the 20 leukemias diagnosed in Woburn since the wells were tested in 1979. Therefore, although we cannot rule out causality by this criterion, neither do we have strong evidence that the temporal pattern was consistent with it.

5. An association exhibits a *dose–response* relationship if the disease rate increases with increasing levels of exposure. Like many environmental agents, "exposure levels" of contaminants in drinking water are vexingly difficult to define and assess. The authors have defined a household annual exposure rate to be the percentage of its annual water supply coming from the contaminated wells. They then cumulated annual exposure rates for each child in their survey and tested for trend by modeling the leukemia rate ratio as an exponential function of cumulative exposure. They found a statistically significant positive trend ($p = .03$), suggesting a dose–response relationship. To reinforce this evidence, it would be useful to examine other functional forms for the trend, such as a step function for high, medium, and low cumulative exposure levels or exposure during individual years when contaminant levels were thought to be higher than others.

6. There is *biological plausibility* to associations between ingested carcinogens and any one of the malignancies collectively called leukemia. The theory that stem cells become malignant after withstanding damage to the genome caused by chemical binding to cellular macromolecules has received ample support from experimental and observational data. Differences in the pathogenesis of the various subtypes of leukemia suggest that they have different etiologies. Thus it would be useful to know the types of leukemia occurring among the 20 cases and to examine whether those exposed to well water were all of one type.

7. The *coherence* of the drinking water-leukemia association measures its agreement with other facts known about leukemias, including their natural history. There is good evidence that in utero exposure to ionizing radiation causes leukemia during childhood. This fact suggests that prenatal exposures to other carcinogens can cause leukemia and, therefore, it supports the present association. Also supportive of causality is the increased risk for leukemia among those with Down's syndrome (Holland, Doll, and Carter 1962), since the Woburn study also found an association between the well water and this birth defect.

8. *Experimental* evidence can add considerable weight to the causal interpretation of an association. Available experimental evidence, however, does not lend credence to a causal role for the contaminant levels found in the Woburn wells in 1979. Among those found, trichloroethylene had the highest concentration (267 ppb), more than 10 times that of other contaminants. But this level is low in comparison with the current occupational standard for trichloroethylene of 100 ppm as an 8-hour time-weighted average. A careful review of the toxicity of trichloroethylene (Kimbrough, Mitchell, and Houk 1985) concluded that on the basis of available evidence from animal, in vitro, and epidemiological studies the risks associated with trace concentrations of this compound in drinking water appear to be minimal or perhaps negligible.

9. If similar associations have proved themselves causal, then by *analogy*, the present one is more likely to be causal. There are no causal associations for leukemia clearly analogous to the present one. Although many might agree that the in utero radiation and leukemia association mentioned previously is causal, they are unlikely to consider it similar to the present one, as the pathogenic mechanisms of physical and chemical carcinogens are dissimilar.

In summary, I rate the leukemia-drinking water association as lacking consistency, specificity, experiment, and analogy, and, therefore, as failing four of Hill's nine criteria for causality. If the association is not causal, it could be due to any one or more of many alternative explanations, some of which are discussed briefly in the article.

The problems we face with toxic wastes indicate that this work will be followed by others like it, and that statisticians will be increasingly involved in analyzing and interpreting these complex data sets. Hill's criteria show that statistical analysis provides input from only one of many disciplines needed to give a complete assessment of the data and a perspective for the difficult issues involved in interpreting the findings.

ADDITIONAL REFERENCES

Gottlieb, M. S., Carr, J. K., and Clarkson, J. R. (1982), "Drinking Water and Cancer in Louisiana," *American Journal of Epidemiology*, 116, 652–667.

Henle, J. (1938), *On Miasmata and Contagie* (translated by G. Rosen), Baltimore, MD: Johns Hopkins University Press.

Hill, A. B. (1965), "The Environment and Disease: Association or Causation?," *Proceedings of the Royal Society of Medicine*, 58, 295–300.

Holland, W. N., Doll, R., and Carter, C. O. (1962), "The Mortality From Leukaemia and Other Cancers Among Patients With Down's Syndrome (Mongols) and Among Their Parents," *British Journal of Cancer*, 16, 178–186.

Kimbrough, R. D., Mitchell, F. L., and Houk, V. N. (1985), "Trichloroethylene: An Update," *Journal of Toxicology and Environmental Health*, 15, 369–383.

Knox, G. (1964), "Epidemiology of Childhood Leukemia in Northumberland and Durham," *British Journal of Preventive and Social Medicine*, 18, 17–24.

Koch, R. (1890), *Uber Bacteriologische Forschung*, in *Proceedings of the 10th International Medical Congress*, Berlin, p. 35.

Pinkel, D., and Nefzger, D. (1959), "Some Epidemiological Features of Childhood Leukemia in the Buffalo, NY Area," *Cancer*, 12, 351–358.

Wilkins, J. R., and Comstock, G. W. (1981), "Source of Drinking Water at Home and Site-Specific Cancer Incidence in Washington County, MD," *American Journal of Epidemiology*, 114, 178–190.

Comment

WALTER J. ROGAN*

Any study of health hazards from neighborhood exposure to hazardous waste sites will produce controversy. The study presented here is in many ways the state of art, and so affords the opportunity to discuss some generic issues as well as those specific to Woburn and the Harvard study.

There have been few studies that have shown a relationship between neighborhood exposure to toxic chemicals and diseases in the residents. Of the various diseases attributed to living near the Love Canal in the late 1970s, only low birth weight (Vianna and Polan 1984) has achieved some kind of acceptance. A well-publicized study of cytogenetic abnormalities (Picciano 1980) was unable to be confirmed (Heath et al. 1984), although the group sampled in the second study was somewhat different from the first. Cancer incidence studies in the area have thus far been negative (Janerich et al. 1981), but insufficient time has elapsed to observe increases in cancers with long latency. In these kinds of studies, even documenting exposure can be difficult. In a North Carolina incident in which polychlorinated biphenyls were spilled on the roadside, absorption of the chemicals was suspected, but body burden, as measured by breast milk levels, was not increased (Rogan, Gladen, McKinney, and Albro 1983). In Times Beach, Missouri, where there was extensive soil contamination by 2,3,7,8-tetrachloro-dibenzo-dioxin (TCDD), bioavailability of TCDD that had been adsorbed to the soil was documented (McConnell et al. 1984), but analytic documentation of exposure has not been published. This is not the place for an exhaustive review of the hazardous waste literature; the point is that the Woburn study is one of the few that has shown some disturbance in health of the neighborhood residents in relation to exposure in a reasonably convincing way.

The controversy usually does not arise in the analysis of the data. We have here the usual problem of model selection. In the absence of knowledge of the mechanism of action of the toxic agent, statistical models are perforce chosen on the basis of convenience, experience, or hope of robustness. This is the best that can be done, and I will not discuss the analysis further.

Data for these studies are of three sorts: exposure, outcome, and nuisance variables or potential confounders. The usual confounders, like the usual suspects, can be identified and rounded up. I can propose no obvious missed confounder here and have only minor quibbles with the way in which confounding was handled. Outcome variables are the illnesses and conditions that the authors hope to attribute to chemical exposures. The major one, childhood leukemia, is ascertained well, and we need have no suspicion that cases either escaped detection or that some cases really have something else. For practical purposes, perinatal deaths are also ascertained fully and remembered well. The other conditions vary a great deal in the confidence we have that the parent's report represents the occurrence of that condition, that the parent's denial means that it did not occur, and that different parent's reports mean the same thing. Amblyopia (impaired vision) and strabismus (squint, sometimes including crossed eyes) are heavily dependent for their reporting on degree to which the parents are concerned. Down's syndrome should be reported well, but the significance of this finding rests on three cases in the high exposed group, and so would appear susceptible to chance. It is not possible to dissect the discovery of blindness; blindness, of course, is apparent relatively soon. Cerebral palsy may not appear until the child is older, in some cases not until school age, and so may reflect some concern on the part of the parents. To some degree, this line of thought can be extended to any of the "soft" interview outcomes, and the distortions produced come under the general rubric of recall bias. The authors address this question to the degree that they can and find no gross evidence of it, and it is of course easy to invoke recall bias here and very difficult to refute it. It is impossible to tell how much the exposed parents knew about their exposure status and whether the results of the water analyses used by the authors played a role in their recall. All we can say is that recall bias may have played a role (i.e., the more the perceived exposure, the better the recall, with events forgotten and not checked among the unexposed), but we do not have strong evidence that it did.

The exposure variable suffers to some degree from the Texas sharpshooter problem. We knew that there were leukemia cases in East Woburn at above expected rates. Thus any exposure variable we choose that is strongly associated with East Woburn will come up positive, an equivalent exercise to painting the target around the pattern of fired bullets. Thus exposure to the East Woburn bank or post office should also show an association with leukemia cases. Comparisons within East Woburn do not suffer from this, but the numbers get small quickly. Water supply to a house is not the best indicator of dose, since cases may well have had substantial, or even the majority, of their water from their school, their work place, or their day-care center. Such an argument does not account for dose/response, except at its extreme. Thus if we accept that the engineering model is reasonable, and we have few details about it to help us decide, then the dose/response relationships are persuasive, whereas the comparisons between East and West Woburn are not. We are left with the problem of exposure to what. The water contained varying amounts of a number of chemicals, none of which is an

* Walter J. Rogan is Medical Officer, Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709.

In the Public Domain
Journal of the American Statistical Association
September 1986, Vol. 81, No. 395, Applications

accepted leukemogen and all of which occur at relatively low concentrations. Toxicologic evaluation is thus very difficult or impossible. Indeed, some of the outcomes, such as cerebral palsy, are perhaps more commonly caused by events in labor and delivery rather than a primary defect in the child.

Biologic or toxicologic plausibility has to enter into the evaluation, although this is a less stringent criterion than some of the others. Any new association will be implausible, simply because it is new, and so we cannot reject unexpected findings out of hand simply because we did not expect them. Nonetheless, we hope that new findings will have some consistency with what we think we already know. The solvents, like trichloroethylene and tetrachloroethylene, are not exotic chemicals. They are produced in huge amounts and used in many occupations and commercial processes, dry cleaning among them. If water contamination in the ppb range is leukemogenic, it is reasonable to expect that workers exposed to high doses over a working lifetime would show high rates. This has thus far not been the case. In the laboratory, on the other hand, these chemicals produced liver tumors in mice and tetrachloroethylene produced leukemia (not the usual childhood leukemia cell type) in rats (National Toxicology Program Technical Reports 243 and 311, in press), albeit at high doses. One can always claim that some combination, perhaps unique to this exposure, adds up to produce leukemia, although assigning an attributable risk to such an ill-characterized exposure is going a bit far. It is also true that the epidemiologic data are not vast, do not include children or pregnant women, and are silent on the outcomes other than cancer and leukemia. The associations thus make us nervous, since they do not fit well with our current concepts of what these sorts of chemicals do at these doses.

Finally, there is the one-of-a-kind problem. In a sense, each of the "dumpsite" exposure problems is different, since there is no standard mix of chemicals, and the vectors by which exposure take place are different. Add to this the problem that such exposed groups are usually relatively small. We may have learned nothing about the human toxicity of any agent from the experience of people in Woburn, which is a shame, but there is an even harder problem. Even in experimental science, the road to consensus and cumulation of knowledge is through replication. In observational studies, particularly in relatively less understood areas like environmental epidemiology, replication is crucial before causality is accepted. The standard example of this is the two dozen or so retrospective and three prospective studies of cigarette smoking and lung cancer that preceded the 1964 Surgeon General's Report. When a phenomenon appears in different places with different designs and different investigators, we are more confident that the associations seen are real. In this case, however, replication is not possible. Thus we cannot know how generalizable the findings are, nor can our usual means of assessing causality, that of replication, help us out.

This study represents what is close to the state of the art in this kind of epidemiology. The problem is certainly real, and our failure thus far to make real sense of either the dumpsite exposure problem or the cancer cluster problem is disappointing. Although we cannot be sure that we will learn a great deal from these investigations, it is clear that they will have to proceed. The care that these investigators took, their willingness to attack a problem with no clear solution, and the vigor with which they have defended their case are laudable. Subsequent generations of epidemiologists and biostatisticians may grin at the crudeness of our current methods for looking at these problems, but the only way to learn how to do it is to do it. These investigators have taken a good early step.

ADDITIONAL REFERENCES

Heath, C. W., Jr., Nadel, M. R., Zack, M. M., Jr., Chen, A. T. L., Bender, M. A., and Preston, R. J. (1984), "Cytogenetic Findings in Persons Living Near the Love Canal," *Journal of the American Medical Association*, 251, 1437–1440.

Janerich, D. T., Burnett, W. S., Feck, G., Hoff, M., Nasca, P., Poladnek, A. P., Greenwald, P., and Vianna, N. (1981), "Cancer Incidence in the Love Canal Area," *Science*, 212, 1404–1407.

McConnell, E. E., Lucier, G. W., Rumbaugh, R. C., Albro, P. W., Harvan, D. J., Hass, J. R., and Harris, M. W. (1984), "Dioxin in Soil: Bioavailability After Ingestion by Rats and Guinea Pigs," *Science*, 223, 1077–1079.

National Toxicology Program (in press), "Carcinogenesis Studies on Trichloroethylene (Without Epichlorhydrin) in F344/N Rats and B63F$_1$ Mice (Gavage Studies)," National Institutes of Health Pub. No. 83-1799, NTP Technical Report 243 (draft).

——— (in press), "Toxicology and Carcinogenesis Studies on Tetrachloroethylene (Perchlorethylene) in F344/N Rats and B63F$_1$ Mice (Inhalation Studies)," National Institutes of Health Pub. No. 86-2567, NTP Technical Report 311 (draft).

Picciano, D. (1980), "Pilot Cytogenetic Study of the Residents Living Near the Love Canal," *Mammalian Chromosome Newsletter*, 21, 86–93.

Rogan, W. J., Gladen, B. C., McKinney, J. D., and Albro, P. W. (1983), "Chromatographic Evidence of Polychlorinated Biphenyl Exposure From a Spill," *Journal of the American Medical Association*, 249, 1057–1058.

Vianna, N. J., and Polan, A. K. (1984), "Incidence of Low Birth Weight Among Love Canal Residents," *Science*, 226, 1217–1219.

Comment

SHANNA H. SWAN and JAMES M. ROBINS*

Episodes of recognized contamination of community drinking water supplies by organic chemicals have become widespread. With rare exceptions, exposure levels found in the work place far exceed those found in the general community, even when chemical spills or leakage from hazardous waste sites contaminate drinking water. For example, wells G and H in East Woburn were found to be contaminated with 267 ppb of trichloroethylene (TCE), the highest concentration of any contaminant found by the Massachusetts Department of Environmental Protection. Yet, by a straightforward toxicological calculation one can demonstrate that the daily molar uptake of a "degreaser operator" exposed for 8 hours at the OSHA permissible exposure limit exceeds the uptake of a Woburn resident drinking 3 liters of water per day from wells G and H by a factor of approximately 1,000.

Therefore, one would expect the risk associated with such community exposures to be small whenever reasonably well designed occupational studies have ruled out large risks. This observation leads to the paradoxical result that "negative" studies on health effects of drinking water contamination may fail to detect important risks, whereas "positive" studies, such as the study of Lagakos, Wessen, and Zelen, tend not to be believed. This follows from the following observations.

Relative increases of 30%–50% in the rates of specific birth defects or leukemia due to widespread contamination of water supplies by organic chemicals would be of considerable public health and community concern especially when there are many communities that are similarly exposed. Unfortunately, because of the nonexperimental nature of epidemiologic studies, whenever an observed relative risk is small, it is nearly always possible to suggest uncontrolled biases or unmeasured risk factors that could explain the observed association. In such situations, epidemiologists are unable to reach consensus on whether the observed association is causal. This lack of consensus will exist even when the study population is of sufficient size to produce a narrow confidence interval for the exposure effect that excludes the null. On the other hand, if the observed relative risk is large (e.g., 3.0 or greater), the observed association is less likely to be entirely explained by sampling variability, confounding, and bias, and consensus on causality is often possible. [It is for these reasons that Hill (1971) chose the observed strength of an association between exposure and disease (implicitly measured in terms of the relative risk) as his cardinal criterion for determining the likelihood that an observed association is causal.] Exceptions to the general rule that large relative risks can lead to consensus occur in epidemiologic investigations of low-level environmental contamination. This reflects the fact that many epidemiologists consider high relative risks to be implausible in such situations based on extrapolation from studies in more highly exposed industrial cohorts. It follows that even though Lagakos et al. find a number of large relative risks, their study cannot, in itself, irrespective of the exact details of design and analysis, result in scientific consensus. Furthermore, given the inherent limitations of epidemiologic investigations, if the true increase in relative risk is of the order of 30%–50%, scientific consensus may never be reached, even when the results are replicated in several well-conducted studies. In fact, it has been argued that epidemiologic studies of low-level environmental contamination should not be performed at all because the plausible excess risks are sufficiently small that they cannot be reliably measured. We believe, however, that a decision to abandon such epidemiologic investigations at this time would be scientifically premature and would fail to consider the potential importance of such studies in guiding those making decisions concerning hazardous waste regulation and cleanup.

Such a decision would be scientifically premature because it is still plausible, although unlikely, that excess risks from environmental exposures may be large enough to be reliably detected by epidemiologic methods. For example, one might argue that it is plausible that the large risk estimates found by Lagakos et al. are causal, as follows:

1. Most occupational studies have been of workers exposed at levels far below those permitted by OSHA. Furthermore, misclassification of exposure in such studies is severe.
2. Occupational studies of pregnancy outcomes are rare, particularly for organic chemicals, and studies of directly exposed children are, of course, nonexistent.
3. The developing fetus, exposed at the critical period, may be at much greater risk than the adult worker.
4. Complex mixtures, such as those found in the Woburn well water, may act synergistically to produce large risks.
5. Levels of contamination may have been much higher before the initial testing of well water in 1979.
6. The dose–response curve for these chemicals may not be linear. In fact, linear extrapolation from occupational studies may underestimate risks at very low doses.

Furthermore, abandoning such studies ignores their potential to assist in decision making. If the true relative risks are in the range 1.0–2.0, decisions concerning the regulation and cleanup of hazardous waste must be made under uncertainty as to the health benefits of such action. As such, the economic cost of regulation and cleanup must be weighed against the likely health benefits, either informally

* Shanna H. Swan is Chief of the Health Assessment and Surveillance Unit, Epidemiological Studies Section, California Department of Health Services, Berkeley, CA 94704. James M. Robins is Assistant Professor of Occupational Health, Harvard School of Public Health, Boston, MA 02115.

Journal of the American Statistical Association
September 1986, Vol. 81, No. 395, Applications

or via formal Bayesian decision analysis. It follows that if the results of these studies can shift the scientific community's beliefs so that small elevations in risk become more believable, this may result in regulation and cleanup that would not have occurred otherwise (Robins, Cullen, and Welch, in press; Robins, Landrigan, Robins, and Fine 1985).

For such a shift in beliefs to occur, many epidemiologic studies must be conducted in areas with similar exposures, they must be meticulously performed to minimize confounding and bias, and evidence must be combined across studies. It is essential to this argument that "hard" end points, such as low birth weight or perinatal deaths, be studied rather than "soft" end points, such as self-reported symptoms that may be systematically biased across all studies. With these thoughts in mind, we will now discuss the specifics of the Woburn study and, in particular, the extent to which Lagakos et al. controlled bias and confounding. We will treat the leukemia study separately from the study of pregnancy outcomes and childhood disorders.

LEUKEMIA STUDY

Prior to the Woburn study by Lagakos et al., the Massachusetts Department of Public Health (MDPH) identified an excess of childhood leukemia, largely confined to East Woburn, for the years 1969–1979. Some possible causal explanations for this finding include the following.

Hypothesis A: General Woburn effect. There is some causal risk factor A, which increased the rate of childhood leukemia in *all* of Woburn, relative to U.S. rates, throughout the study period. (This unknown risk factor may be environmental, occupational, genetic, etc.)

Hypothesis B: East Woburn effect. There is an unknown causal risk factor B, which increased the rate of childhood leukemia in all of East Woburn relative to West Woburn throughout the study period.

Hypothesis C: Well effect. Contamination in wells G and H increased the rate of childhood leukemia among individuals exposed to well water before the wells were closed in June 1979.

Although numerous other hypotheses might be listed, we limit ourselves to these for the purpose of this argument. Note that these hypotheses need not be mutually exclusive. Furthermore, all might be false. That is, the observed cluster may have arisen by chance.

Before the recent study of Lagakos et al., the MDPH could not address Hypothesis C, the well effect, because exposure data were not available. The current study by Lagakos et al. evaluates Hypothesis C by using recent information on the space–time distribution of water from wells G and H and on eight additional leukemia cases.

We will now consider the evidence from Lagakos et al. We will discuss how this new evidence might alter various prior beliefs about the aforementioned causal hypotheses. By "prior beliefs" we refer to beliefs that may have been held by scientists prior to reviewing the results of the Lagakos et al. study but after the results of the MDPH study became available. We feel that this exercise is informative even if most epidemiologists feel unable to quantify their prior beliefs. We examine Lagakos et al.'s results on leukemia in two stages to reflect the authors' analysis.

We first consider the data on exposure, year of birth, and residential history for individuals born prior to 1980 (the pre-1980 birth cohort). Only these data contributed to the score test based on the failure-time regression model used by Lagakos et al., as follows:

$$h\{t \mid x(t), y\} = h_y(t) \exp\{\alpha x(t)\}. \quad (1)$$

Note that the authors allowed for a well effect, represented by the coefficient α, and a general Woburn effect, represented by the background hazard $h_y(t)$ [since they did *not* constrain $h_y(t)$ to be equal to U.S. age-specific leukemia rates]. The authors effectively eliminated an East Woburn effect, a priori, by failing to include in Equation (1) a proxy for the unmeasured risk factor described under Hypothesis B. This proxy could have been included by adding a multiplicative term for an East Woburn effect in the hazard function [a possible form for which is $\exp\{\beta c(t)\}$, with $c(t)$ being cumulative years lived in East Woburn]. Under this prior assumption of no East Woburn effect, the authors find that the likelihood of the observed data is maximized if Hypotheses A and C are both true, since the estimate of the well effect is significantly different from zero; yet the well effect does not explain the entire leukemia excess. This might have been predicted since, given that wells G and H served only East Woburn and the MDPH study had found an East Woburn leukemia cluster, it is possible that almost any distribution of water within East Woburn would have produced a significant well effect when fitting Equation (1).

In Table 8, the authors do attempt to control for possible confounding by an unmeasured risk factor in East Woburn, but the test reported in that table has poor power and is not restricted to the pre-1980 birth cohort. Had they included a covariate for an East Woburn effect in (1), it is quite possible that neither the "well coefficient" nor the "East Woburn coefficient" would have been significant after controlling for the other because of the high correlation between cumulative years of residence in East Woburn and cumulative exposure to wells G and H for the pre-1980 birth cohort. In fact, it is possible that the estimate of α controlling for an East Woburn effect would be negative. This could occur if, for example, within East Woburn, the leukemia cases were clustered in areas and years other than those maximally served by wells G and H. If so, the data from the pre-1980 birth cohort could result in a decrease in the relative odds of Hypothesis C compared with Hypothesis B.

Even if coefficients for both the East Woburn and the well effect were positive, the posterior odds for a well effect could be diminished if the estimate of the East Woburn coefficient was greater in magnitude and had a more extreme p value than that of the well coefficient and one believed a priori that it was quite unlikely that there were two distinct causes for the increased rate of leukemia in East Woburn.

Belief in "one cause" would rest on the assumption that it is highly unlikely, a priori, to find two leukemogenic factors that were operating at the same time and in the same place. If for most leukemogens, however, the expected elevation in risk is small, it is not clear that the Woburn cluster would be more easily explained by a single leukemogen (plus sampling variability) than by two leukemogens.

We now consider the additional effect of the leukemia data for the post-1980 birth cohort on our causal inferences. The excess of leukemia in this cohort occurred in West Woburn, with no cases in the most exposed areas of East Woburn (Zones A–C). These data support Hypothesis C (well effect) compared with Hypothesis B (East Woburn effect), since Hypothesis C predicts an end to the East Woburn excess after the wells are closed. [It follows that had an East Woburn term been included in Equation (1) and had this equation been fit to the entire data set (both pre-1980 and post-1980 birth cohorts), it is likely that the estimate of α would have been positive even had this estimate been negative in the analysis restricted to the pre-1980 birth cohort.] The new cluster in West Woburn, however, is evidence for Hypothesis A. In fact, the post-1980 birth cohort data could diminish one's belief in the plausibility of a well effect, if one believed that in Woburn as a whole there was at most one cause of excess leukemia.

Under the assumption that there was no confounding by an unmeasured risk factor in East Woburn, can the observed exposure effect be explained by selection and/or misclassification bias? Ascertainment of leukemia cases appears to be complete. Furthermore, it is likely that the residential histories of the leukemia cases are correct. In addition, the authors use the known residential distribution of all of Woburn to compute expected cumulative exposures in one of their analyses. This choice eliminates the possibility of selection bias. Misclassification bias can arise only from the assignment of exposure to residential areas. Any such misclassification would be nondifferential with respect to outcome and, therefore, must lead to bias toward the null. Thus we feel that these biases were unlikely to explain the positive result reported by Lagakos et al. On the other hand, Lagakos et al. chose to investigate the incidence of childhood leukemia because it was the sole cancer found by the MDPH to be elevated in East Woburn. The additional knowledge that rates of no other types of cancer were elevated in East Woburn would, in general, have the effect of diminishing one's belief in a well effect on childhood leukemia unless one believed a priori that either childhood leukemia would be the cancer most likely to be elevated (e.g., because of short latency) or the suspect carcinogen causes only one type of cancer.

In summary, Lagakos et al. should ideally have searched for a well effect while controlling for an East Woburn effect. If neither the well nor the East Woburn effect were significant when controlling for the other because of a high correlation between residence in East Woburn and exposure to wells G and H, the authors' study would add little evidential weight to the well hypothesis, above and beyond that of the MDPH study. Taken as a whole, the data presented by Lagakos et al. could either increase or decrease one's belief that wells G and H were causing leukemia in East Woburn, depending on the particular prior correlations that one believed existed between Hypotheses A, B, and C (e.g., depending on whether one believed that at most one of these hypotheses could be true) and on whether the estimate of α controlling for an East Woburn effect was positive or negative in the pre-1980 birth cohort.

STUDY ON PREGNANCY OUTCOMES AND CHILDHOOD DISORDERS

We now consider the results on pregnancy outcome and childhood disorders. In this phase of the study, outcomes were determined through interview. Since the individuals interviewed knew about the purported health hazards associated with the exposures to wells G and H, respondents living in East Woburn might be more likely to recall real or imagined health events than residents of West Woburn, because of their concern over the possible adverse health effects of wells G and H. Furthermore, because many interviewers were also community residents and the blinding of interviewers was fallible, it was possible that there was interviewer bias as well as recall bias by respondents. The following discussion of recall bias applies equally to interviewer bias.

The authors suggest that recall bias would be unlikely to explain the observed associations. First, they show that in the years the wells were not pumping (1960–1963, 1973, 1980–1982), the rates of adverse pregnancy outcomes were similar in East and West Woburn (Table 8). This observation does not offer convincing evidence, however, against the hypothesis that apparent exposure effects were due to recall bias, for the following reasons. First, East Woburn residents were aware that the wells had been closed in 1979. Thus after 1980 their concern about adverse pregnancy outcomes may have abated. Second, children born in the years 1960–1963 would be 19 years or older at the time of interview. It is quite possible that parents of such grown children were less concerned about the possible adverse effects of the wells on pregnancy outcome than parents of younger children. (The potential for recall bias is generally considered to increase with increasing period of recall. Here we are suggesting that this increased potential for recall bias may be outweighed by the diminished concern of parents of grown children.) Thus the year most relevant to the authors' argument is 1973. The authors do not present disaggregated data for 1973, and it is likely that there were not sufficient pregnancies in this one year to address this bias. The authors also compare the rates of perinatal deaths, ear and eye anomalies, and CNS, chromosomal, and oral cleft anomalies within East Woburn for the years during which wells G and H were pumping to years when these wells were not pumping (Table 9). But as argued previously, this is not a convincing test of the absence of recall bias, since in the years 1980–1982, when the wells were not pumping, concern over health effects may have diminished because of closure of the wells in 1979. Nonetheless,

Tables 8 and 9 do offer some, even if not strong, evidence against recall bias.

An alternative method of addressing this question of recall bias is to determine whether the apparent exposure effect persists when the analysis is restricted to East Woburn only. It would have been desirable to carry out this test including residents of East Woburn with no exposure to wells G and H. Since these data were not available, we reanalyzed the data from Table 5 for individuals with exposure scores of .01 or greater. We found that only for premature deaths in 1970–1982 and cardiovascular defects in 1960–1982 was there a significant trend in rates ($p < .05$). In the latter case, the trend was in the direction of decreased risk with increasing exposure ($p = .01$). Thus, by this test, there is not much evidence for a dose response above and beyond an East Woburn effect.

Even if we had discovered a persistent well effect among East Woburn residents, could this be explained by recall bias? One might think not, since prior to the MDPH water distribution study, no East Woburn resident knew what fraction of their water had been supplied by wells G and H. The issue is not, however, whether people had factual information about the proportion of their water that was obtained from these wells but rather whether their beliefs as to their water source correlated with reality. Residents might well have believed that living further east within East Woburn was associated with increased contamination of drinking water. A look at Table 1 and Figure 1 suggests that such beliefs would not have been far wrong. One approach to controlling for recall bias would have been to ask respondents to quantify their beliefs as to the temporal pattern of exposure to contaminated drinking water in their area of Woburn, compared with that in other areas. Then, in the analysis, one could have stratified on perceived exposure when estimating the effects of true exposure. (This method of controlling for recall bias should be successful unless people consciously lie about their perception of exposure.)

As a final check on recall bias, the authors attempted to obtain medical verification for a stratified sample of 96 disorders. They were able to obtain medical records for 66 of the 96, of which 62 were confirmed. Among the 66 cases for which medical records were available, however, the authors did not find an important differential false-positive rate between exposed and unexposed areas. Such bias has been reported, however, in other studies. Recently a study of health effects of a hazardous waste site in Riverside, California [the "Stringfellow" study (1986)], which included more complete verification of adverse pregnancy outcomes, did find differential recall by study area. In the areas most likely to have been exposed in this study, two-thirds of reported spontaneous abortions and low birth weights were in error, whereas none of these reported outcomes were found to be in error in the control area. In contrast, in a study of reproductive outcomes in Santa Clara County, California, in a community exposed to fairly high levels of organic solvents in drinking water [the "Fairchild" study, see *California Department of Health Services* (1985)], false-positive rates were similar in exposed and unexposed subjects. Based on our results, it appears to be important that medical records be obtained to verify all adverse outcomes reported on interview.

Lagakos et al. did not verify negative histories. In the Fairchild study, the false-negative rate was evaluated by reviewing the medical records of a 50% sample of women reporting no adverse reproductive outcomes. This error rate was found to be neglible. This is reassuring, since it suggests that in interview studies of adverse reproductive outcomes in relation to environmental contaminants, negative responses may not need to be verified. But there is an important caveat here. The Fairchild study included as birth defects only those conditions that were considered "reportable" by the California Birth Defects Monitoring Program. These defects are serious, verifiable through hospital records, and reported uniformly. They exclude such conditions as unqualified "heart murmur" and "crossed eyes," both of which were included as birth defects by Lagakos et al. Reportable defects are less likely to be forgotten and underreported. As such, the false-negative rate for these defects should be low, as was found in the Fairchild study. For the study of Lagakos et al., however, which included outcomes that are less likely to be reproducible, we have no assurance that false-negative rates would be equally reassuring. This comparison is further confused by the fact that women were interviewed within 2 years of their pregnancy in the Fairchild study. Since pregnancies were recalled over almost 20 years in the Lagakos et al. study, as in the Stringfellow study, reporting errors are more likely. Moreover, reporting errors may well have been differential with respect to exposure at Woburn, as they were at Stringfellow.

The possibility for recall bias for childhood disorders is, if anything, even greater than for pregnancy outcomes because these end points are "softer." As the previous discussion makes clear, we remain unconvinced that bias was not a problem for any of these end points, although we have no evidence that such bias was present. If one believed a priori that significant bias is more likely than a large well effect, the study of Lagakos et al. probably would not change one's beliefs. In addition, even if, as in the leukemia study, no recall, interview, or selection biases were operating here, the possibility of confounding by unmeasured risk factors still remains.

DISCUSSION

What are the implications of the previous discussion for environmental epidemiology?

To change beliefs of the scientific community as to the likelihood of significant public health consequences when expected relative risks are small, evidence from many studies must be combined. Even though we can never rule out positive confounding by unmeasured causal risk factors in any single study, it may be reasonable to believe that after controlling for measured risk factors, the uncontrolled factors are not systematically associated with exposure across studies. Thus by combining evidence from multiple studies,

bias due to unmeasured risk factors may be eliminated. To obtain adequate power one would need to ensure that the exposures under study are truly similar and that individuals themselves are correctly classified with respect to exposure. This calls for a level of exposure assessment that is seldom achieved in environmental epidemiology, Lagakos et al.'s study being an exception. Accurate exposure assessment allows for internal (i.e., within-community) comparisons as in the Lagakos et al. study. Such internal comparisons should lead to increased power (and often to decreased bias), because (a) the variation in unmeasured risk factors, when controlling for measured risk factors, should be less within than between communities and (b) ranking of individuals by exposure need only be accurate within communities. We should add that this exposure assessment has two components, one environmental and one individual. In the Fairchild study we obtained information on the amount of tap water consumed during pregnancy. This covariate was seen to vary widely in this population.

Furthermore, any bias, such as recall, interviewer, or selection bias, that can systematically distort the results of each study in the same direction must be prevented. Because of the real potential for recall bias documented in the Stringfellow study, outcome measures must be those recorded by physicians and, for the less serious outcomes, must have been recorded in medical charts prior to any community knowledge of exposure to toxic chemicals. This is necessary because individuals with less severe outcomes might eventually seek medical care based on perceived exposure. In contrast, serious outcomes, such as leukemia, will always result in medical care. For outcomes such as cancer and birth defects, registries, where available, will allow studies to be performed relatively easily. For outcomes such as spontaneous abortions, one needs to examine individual hospital records, since registries are not available. Moreover, to minimize interviewer bias and selection bias, professional interviewers blind to exposure should be used. Nonreponse should be minimized; for example, the higher participation rate obtained at Fairchild required up to 10 repeat calls per interview compared with 3 reported by Lagakos et al. Finally, all positive outcomes, as well as a sample of negative outcomes, should be verified. Since the false-negative rate will be low for rare outcomes, this sample must be a large fraction of the total population. This is an expensive undertaking.

It may be that the relative risks due to particular low-level contaminants are much greater for subclinical biochemical and physiologic end points (e.g., DNA adducts, T-cell function) than for cancer and birth defects. If such outcomes can be shown to predict clinical disease, epidemiological studies focused on these end points would be less sensitive to the issues of bias, confounding, and chance discussed previously. These studies would also be expensive, and this branch of epidemiology, although promising, is still in its infancy.

Is expenditure of the kind of money necessary to carry out the projects described above reasonable and realistic? Large amounts of money have already been spent, but much of it has been wasted. For example, industry has spent a great deal of money critiquing the results of studies of Woburn and Love Canal. Many of the shortcomings of these studies, however, are the result of inadequate funding. For example, in Woburn, volunteer interviewers were used, positive interview responses were not adequately followed up, and negative interview responses were not followed up at all. If the industries involved had initially supported independent investigators to carry out careful studies, everyone would have benefited. Because of the controversial nature of these studies, we believe that an advisory committee, on which the community, industry, and the government are represented, is essential to this process. In California we have used such committees with a good deal of success (e.g., at Stringfellow and Fairchild). Should the government itself undertake to fund such studies? One argument against government funding of these studies rests on the view that it is unlikely that they will produce results that will lead to consensus concerning causality. But, as we argued previously, such studies are important for regulatory decision making if they can influence the beliefs of the scientific community, even though they cannot lead to consensus. Furthermore, while large uncertainty as to the magnitude of the health effects of environmental contamination remains, many thousands of citizen hours are spent each year in trying to document suspected health effects, trying to get public officials to conduct studies, and lobbying for cleanup. These hours must be considered in any cost-benefit analysis. Such an analysis may well demonstrate that it is cost effective to clarify the issues by performing meticulous studies of areas such as Woburn. These comments should not be interpreted as an argument in favor of postponing cleanup or diverting funds from cleanup activities to research.

The epidemiological community, steeped in the Hill criteria for causality, often errs in the direction of regarding associations that fail to meet these criteria as noncausal. From a public health perspective, it would seem prudent to err in the direction of treating such associations as causal, until such time as further evidence suggests that the risks are trivial. We believe that the study by Lagakos et al. will have the effect of reorienting debate within the scientific community over the health effects of low-level contamination to a public health perspective. Given the level of risk reported by Lagakos et al. it is incumbent upon industry or government to show that these findings are the result of bias, by making resources available to conduct the necessary studies. Prior to recent studies such as that of Lagakos et al. or Fairchild, it was argued that no data existed linking low-level pollution to health outcomes. Yet adequate funds were not made available to obtain such data. Therefore, it hardly seems fair to cry "foul" when, because of limited funds, Lagakos et al. performed a study that failed to meet the most rigorous epidemiological standards.

We commend Lagakos et al. for undertaking a difficult study with limited resources in a highly charged political environment. Many of the problems we have described here were inevitable given the limitations of resources. To the extent that this study has sparked debate and brought

the attention of the scientific community to the problem of documenting the adverse health effects of low-level environmental contamination, the authors have done a service. We hope that their initial work will be followed up by well-funded, carefully designed studies of Woburn and other communities exposed to low-level, environmental contamination.

ADDITIONAL REFERENCES

California Department of Health Services (1985), *Pregnancy Outcomes in Santa Clara County, 1980–1982: Reports of Two Epidemiological Studies,* State of California Publications Section, Pub. 7540-958-1301-5.

Hill, Bradford (1971), *Principles of Medical Statistics* (9th ed.), London: Lancet Ltd., p. 245.

Robins, J. M., Cullen, M. R., and Welch, L. S. (in press), "Improved Methods for Discerning Health Impacts of Current Technologies," in *Environmental Impacts on Human Health: An Agenda for Long-Term Research and Development,* ed. Sidney Dragen, New York: Praeger.

Robins, J. M., Landrigan, P. J., Robins, T. G., and Fine, L. (1985), "Decision Making Under Certainty in the Setting of Environmental Health Regulation," *Journal of Public Health Policy,* 6, 322–328.

Stringfellow Health Effects Study: An Epidemiological Health Survey of Residents of Glen Avon and Rubidoux, California (1986), Report prepared by Dean Baker and Sander Greenland for the California Department of Health Services, Los Angeles.

Comment

ALICE S. WHITTEMORE*

The authors of this interesting article have attacked a difficult problem with limited funds, and in a highly politicized climate. Their findings are disturbing. Perhaps most worrisome, children from households served by the contaminated wells had significantly greater leukemia rates than did children from other households. This association will strengthen the lawsuit brought against two major corporations by the families of seven children, victims of leukemia that their parents say was caused by industrial pollution of drinking water.

But is the association causal?

The authors are properly careful to warn us against leaping to that conclusion. Their analysis and discussion, however, sidestep the causality issue and fall short of providing the framework we need to wrestle with it.

The struggle to distinguish causal from noncausal relationships predates the disciplines of epidemiology and biostatistics. In 1840, Jakob Henle published postulates for evaluating a causal relationship between a new infectious agent and a clinical disease [see Henle (1938) for an English translation]. His pupil, Robert Koch, developed these postulates and presented them in 1890 before the International Medical Congress in Berlin (Koch 1890). More recently, Sir Austin Bradford Hill (1965) modified them for epidemiological studies of environmental agents and noninfectious diseases. Applied to the present problem, Hill's postulates require that the association between contaminants in drinking water and the leukemias should have (1) strength, (2) consistency, (3) specificity, (4) temporality, (5) a dose–response relationship, (6) biological plausibility, (7) coherence, (8) experiment, and (9) analogy. It is important to assess the drinking water-leukemia association, according to each of these criteria.

1. The *strength* of the association describes the magnitude of the disease rate in the exposed versus the unexposed. In Section 4.1 we calculate that children from households served by the wells had $\exp(1.11) = 3.03$ times the rate of leukemia incidence than did other Woburn children in the survey. The article provides us with no confidence limits for this estimate, so it is difficult to interpret it as a measure of strength. As it stands, it is moderately strong: weaker than the lung cancer rate ratios of 10 or more experienced by lifelong heavy smokers relative to nonsmokers, and stronger than the heart disease rate ratios of two or so associated with smoking. The authors remind us that if a contaminant in the water did cause some of the leukemias, the large error with which they measured exposure to that contaminant produces a bias toward unity in the observed rate ratio.

2. The article does not tell us much about the *consistency* of the leukemia-drinking water association in relation to other epidemiological studies. It is provocative that unexplained clusters of childhood leukemias have been found in many parts of the world (e.g., Knox 1964; Pinkel and Nefzger 1959). Could it be that unmeasured contaminants in the drinking water caused them? The evidence from other data does not support such a conclusion for the present study. At least two investigations of contaminants in drinking water and site-specific cancers found no association with adult-onset leukemia (Gottlieb, Carr, and Clarkson 1982; Wilkins and Comstock 1981). Thus the current findings are not consistent with the results of other studies.

3. The *specificity* of an association is a measure of its uniqueness. For example, nothing other than exposure to polyvinyl chloride monomer has been associated with angiosarcoma of the liver, and conversely, this compound has not been strongly associated with other diseases. By contrast, exposures to benzene, ionizing radiation, and certain viruses have been associated with one or more of the adult and childhood leukemias, and the authors tell us that other

* Alice S. Whittemore is Professor (Research), Department of Family, Community and Preventive Medicine, Stanford University School of Medicine, Stanford, CA 94305.