An Analysis of Contaminated Well Water and Health Effects in Woburn, Massachusetts: Rejoinder

S. W. Lagakos; B. J. Wessen; M. Zelen

Journal of the American Statistical Association, Volume 81, Issue 395 (Sep., 1986), 611-614.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28198609%2981%3A395%3C611%3AAAOCWW%3E2.0.CO%3B2-B

Rejoinder

S. W. LAGAKOS, B. J. WESSEN, and M. ZELEN

We thank the discussants for their remarks, which give new insights about the study in Woburn and raise broader issues about community health studies of environmental contamination. Such discussion is useful because an increasing number of similar types of studies are being conducted and the scientific community has much to learn in terms of their design and analysis.

Because of limitations of space, we cannot respond in detail to all of the points that have been raised. Instead we shall focus on those we feel to be the most important, namely, technical problems of analysis (multiplicity of testing, use of asymptotic theory, sensitivity issues), recall and other bias, presence of a regional confounder, causality, and some broader issues. We have organized our response within these topics.

1. TECHNICAL PROBLEMS OF ANALYSIS

Prentice notes that the normal approximation used to evaluate the test statistic for the leukemia analyses might be inaccurate, citing the radiation example studied by Hoel and Jennrich (1984). We find this illuminating, especially since this type of inaccuracy would not be expected based on the simulation results of Johnson, Tolley, Bryson, and Goldman (1982) (although their focus was on tests based on the maximum likelihood estimator, not the score test). Additional investigations of the small sample behavior of the score test would be useful, especially for situations in which the covariate has a skewed distribution, as is sometimes the case for time-dependent covariates. Prentice also illustrates that the score test from Cox's proportional hazards model can be heavily influenced by a single observation in studies with small samples. We return to this point later in our discussion.

Several discussants comment on the issue of interpreting studies in which multiple tests have been carried out. A total of 25 end points for association with G and H exposure were tested. Six of these were judged to be significant. Thus if there were no real associations between these end points and G and H exposure, we might expect one or two positive associations to occur purely by chance. Prentice assesses the multiplicity issue somewhat differently by looking separately at the 8 adverse pregnancy outcomes, the 6 (not 12) time period interactions for those adverse pregnancy outcomes having nonsignificant main effects, and the 9 childhood disorders. MacMahon raises the multiplicity issue with respect to congenital anomalies and asks how one should interpret the finding of two or three significant associations with wells G and H when "several hundred categories were or could have been tested." What needs to be remembered for the question of multiplicity is not the number of distinct anomalies that are present in the data but rather the number of *groups* of anomalies that were tested, the point being that a heterogeneous group of anomalies is no more likely to produce a spurious association than a homogeneous group. For congenital anomalies, we made eight tests, not hundreds.

MacMahon criticizes our classification of cerebral palsy, mental retardation, and certain other conditions as birth anomalies. He also questions the grouping of distinct types of anomalies into broader categories. With respect to the first point, our inclusion of these categories is not new and was motivated by Warkany (1971), who stated "the area of cerebral palsy overlaps, in part, that of mental retardation, and both areas are related to congenital malformations." With respect to the other types of conditions considered as anomalies, we note that there are widely differing views in the published literature regarding which specific conditions should be considered as birth anomalies (see Christianson, van den Berg, Milkovich, and Oechsli 1981; Myrianthropoulos and Chung 1974). Our definitions more closely followed Myrianthropoulos and Chung (1974).

Next consider our grouping of anomalies into the "eye/ear" and "CNS/chromosomal/oral cleft" categories. We first point out to MacMahon that we did not "invent" the eye/ear category—it was used by the National Collaborative Perinatal Project (Heionen et al. 1983), one of the largest and most definitive studies of birth anomalies conducted in this country. The rationale for our grouping of CNS, chromosomal, and oral cleft anomalies is detailed in our article and is based not on any claim of biologic similarity of these end points but rather in accordance with an elevated prior probability of an association. The hope was to obtain greater sensitivity, albeit at the expense of reduced specificity. Two alternatives to our strategy of forming a few groups are (a) to test individually each of the 76 specific birth defects listed in Table 3 or (b) to form a single category for all birth defects. The former strategy leads to severe problems of multiplicity and to a number of individual tests with little or no statistical power. The latter strategy avoids the multiplicity and "small numbers" problems but increases the risk of missing an important association between exposure and some subcategories of anomalies, as illustrated by the thalidomide example cited in our article.

The issue of how to group anomalies for purposes of analysis in a small or medium-sized study is a complicated one, and we had hoped that the discussants would have commented on it. MacMahon makes clear his unhappiness with our grouping of eye and ear anomalies and CNS, chromosomal, and oral cleft anomalies but stops short of describing and evaluating the merits of other options.

MacMahon objects to our inclusion of certain types of eye anomalies in the eye/ear category and seems to inter-

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

pret the disproportionate number of exposed cases among these as evidence of bias or, worse yet, "gerrymandering." Because the eye/ear category was positively associated with G and H exposure, however, it inevitably follows that there will be a disproportionate number of exposed cases in this category. Indeed, this is evident in both the subset of cases that he refers to (4 of 7) and for the remaining cases (5 of 11). Thus we cannot see the logic of his point.

In response to some other points raised by Prentice and Whittemore, (a) the Massachusetts Tumor Registry uses the SEER criteria for classifying all cancers, (b) the three nonacute lymphocytic leukemia cases were subjects 1 (AML), 9 (AML), and 13 (AML/CML), and (c) we regard the use of two-sided tests to be inappropriate for hypotheses concerning G and H exposure.

2. RECALL AND OTHER BIAS

We and most of the discussants have considered the possibility that recall bias might have had an effect on the associations with wells G and H. Swan and Robins point out that additional medical confirmation of the positive end points and medical confirmation for respondents reporting no adverse health outcomes would be an effective way of detecting and reducing the influence of such biases. We agree and add that even this suggestion is not foolproof, because a differential tendency among exposed and unexposed individuals to seek medical attention could also cause recall bias and would not be detectable by examination of medical records.

In our view, the key questions for empirically assessing the possibility of recall bias are as follows: (a) Is there any evidence of recall bias in the data? and (b) How convincing is the evidence that recall bias is negligible? Neither we nor the discussants claim to have found any evidence that recall bias is present, but we clearly have different views about the chances that it is present and about what could/should be done to check and control for it.

The empirical evidence that argues against recall bias having a material effect on the main results of the study are the following: (a) the small and similar proportions of "false positives" among reported cases from East and West Woburn; (b) the similarity in rates for *unexposed* persons from East and West Woburn, for those end points that were associated with exposure to wells G and H; (c) evidence of a dose–response relationship within East Woburn; (d) the similarity of East and West Woburn with respect to the rates of spontaneous abortions (the most common of the adverse pregnancy outcomes and probably more susceptible to recall bias than perinatal deaths and most birth anomalies); and (e) the similarity between the rates of those end points not associated with G and H exposure and state/national rates, for those conditions for which adequate information was available for a comparison.

It is interesting to remind ourselves how these opportunities to check for recall bias compare with opportunities in other observational studies. In case-control studies, cases and controls are asked to recall past "exposures" (e.g., information about diet, medications, exercise, lifestyle, etc.). The concern for recall bias in these studies is based on the possibility that cases, having thought about what might have caused their disease, are more likely to remember and/or report exposures than controls. The survey we conducted requested information from individuals about reproductive and childhood anomalies, which allows the possibility to verify respondents' reports by examining medical records. In contrast, since many of the exposures solicited in case-control studies cannot be independently confirmed, the opportunity to check for recall bias by confirmation often does not exist. Similarly, the analog of our comparison of unexposed persons from East and West Woburn does not seem to exist in case-control studies. This comparison with case-control studies is not intended to suggest that recall bias is absent in the present study, but to remind us that, as Rogan notes, the possibility of recall bias can always be invoked, but never ruled out.

MacMahon comments on the "imperfection" of the survey by noting a poor response rate. We note that the nonresponse rate of 18% is better than most phone or mail surveys and was uniform across Woburn. It is also important to remember that the survey completed interviews for 55% of the target population. In practice, it is unusual to have such a high sampling fraction, and this limits the opportunity for nonrepresentative samples.

3. REGIONAL CONFOUNDER

There has been considerable discussion of whether the associations with wells G and H could have been due to some factor whose distribution was "regional," that is, more prevalent in the eastern part of Woburn, where wells G and H delivered water, but with no temporal component. In the article we address this issue in two ways. First, we note that the baseline rates (i.e., rates among persons not exposed to wells G and H) in East Woburn are similar to those in West Woburn (Table 9). If the association with wells G and H were the result of a regional confounder, we would have expected these East Woburn rates to be higher when the wells were not in service. Second, we show that within East Woburn, the rates of perinatal deaths (post-1970), eye/ear, and CNS/chromosomal/oral cleft anomalies are higher in the years wells G and H were pumping than in the years they were not pumping (Table 9). This would not have been expected if the G and H association were due to a regional confounder. Another way to examine the latter point is to investigate the existence of a dose–response relationship within East Woburn. Swan and Robins begin to address this point by examining the data in Table 6 for exposed persons (who, by definition, are from East Woburn). But they omit the unexposed data for East Woburn, part of which could have been culled from Table 8. The inclusion of these unexposed data produces results similar to those produced by the full analysis. The rates (per 1,000), based on all of the unexposed pregnancies from East Woburn, are as follows: perinatal deaths, post-1970—8 (4/520), musculoskeletal anomalies—14 (11/799), cardiovascular anomalies—8 (6/799), eye/ear anomalies—4 (3/799), CNS/chromosomal/oral cleft anomalies—8 (6/799), and "other" anomalies—13 (10/799). When combined with the results for exposed pregnancies in Table 6, this information gives the raw data for all of East Woburn. We repeated the logistic regression analyses for association between G and H exposure and these end points by controlling for the same risk factors as in the original analyses but using only the data from East Woburn. This gave the following: perinatal deaths, post-1970—$\hat{\alpha} = 2.34$, $P = .006$; musculoskeletal anomalies—$\hat{\alpha} = -.72$, $P = .78$; cardiovascular anomalies—$\hat{\alpha} = -1.49$, $P = .80$; eye/ear anomalies—$\hat{\alpha} = 2.08$, $P = .005$; CNS/chromosomal/oral cleft anomalies—$\hat{\alpha} = .99$, $P = .10$; and "other" anomalies—$\hat{\alpha} = -.24$, $P = .60$. Similar reanalyses of the leukemia, lung, and kidney associations, based on Cox's regression model, gave the following: leukemia, cumulative exposure—$\hat{\alpha} = .21$, $P = .16$; lung—$\hat{\alpha} = .11$, $P = .13$; and kidney—$\hat{\alpha} = .25$, $P = .08$. If the significant associations in our original analyses were caused by a regional confounder, one would not expect to see any evidence for a dose–response within East Woburn for these end points. The larger significance levels for the Cox regression tests are in part due to the fact that very few subjects have no G and H experience.

Swan and Robins suggest another type of analysis of childhood leukemia, based on fitting a proportional hazards regression model with both a "well" and an "East Woburn" term. They describe how to modify prior beliefs of a "well" hypothesis after fitting this model. Although we find this approach interesting, we question its appropriateness in the present problem. Swan and Robins seem to presume that the "East" hypothesis had been raised by the MDPH/CDC study and that the "well" hypothesis was raised only afterwards. But just the opposite occurred. The MDPH/CDC study was initiated largely because of the discovery that wells G and H were contaminated. Because of a lack of quantitative estimates of the distribution of water from wells G and H, the MDPH/CDC was unable to test formally for an association between G and H exposure and childhood leukemia. Swan and Robins seem to equate the time at which we formally tested for a G and H association with the time the G and H hypothesis was raised. On this premise they propose to assess the plausibility of the "well" hypothesis, in part by determining if a significant dose–response relationship exists after controlling for an "East" effect. We disagree with this way of appraising the "well" hypothesis. It is important to distinguish between an identified factor that is associated with risk and a post-hoc geographic description of where an elevated rate was found. The approach proposed by Swan and Robins is also complicated by the fact that they seem to define "East" Woburn in the same way that we do. Our definition was determined entirely by the estimates of G and H exposure ("West" Woburn was defined to be the portion of the town estimated never to have received any water from wells G and H, and "East" was its complement). The use of this definition of "East" in the Swan–Robins proposal induces collinearity between the "East" and "well" variables and thereby complicates the interpretation of the results of their analysis.

Prentice raises the possibility of repeating the leukemia analysis of G and H exposure with stratification by both year of birth and census tract. Since three of Woburn's six census tracts appear to lie entirely in West Woburn (as we define "West"), this test would use only the data from Woburn's three other census tracts. We were able to compute the contribution to this test from census tract 3334, for which the data were computerized, and obtained $\hat{\alpha} = .50$, $P = .06$ for the cumulative metric and $\hat{\alpha} = 1.13$, $P = .07$ for the none–some metric. In discussing this test, Prentice discusses the possibility of a myriad of other factors that distinguish East and West Woburn. To our knowledge, no factor has been identified that distinguished East and West Woburn except G and H exposure, and this possibility was raised before the MDPH study.

Several other types of statistical information were used to assess the possible causality of the association between wells G and H and leukemia. MacMahon estimates a "risk ratio" (RR) for both the "cumulative" and "none versus some" metrics of G and H exposure, and on the basis of the latter being larger, suggests that "the underlying relationship is markedly nonlinear." We are not entirely clear about his point but note that his estimate of 2.02 is the estimated RR for an exposed person with average exposure and not the average RR among exposed persons. That the former always will be less than the latter follows from the convexity of the estimate for RR ($=\exp\{.33^*x(t)\}$). What is more, the skewness of the distribution of cumulative exposures suggests to us that his estimate may be considerably less than the average RR that we think he is trying to estimate.

MacMahon also discusses our "cruder measure" of exposure (see Sec. 5.3) in connection with his interpretation of the leukemia results. As we indicate in the article, however, this crude measure was used only to reanalyze the pregnancy outcome data—it was not applied to either the leukemia or childhood disorder data.

Swan and Robins and MacMahon discuss the fact that the G and H association with leukemia does not explain all of Woburn's leukemia excess, relative to national rates. We wish that they would have discussed this issue in greater detail. For example, it would seem that a fuller interpretation of the result would be possible if one took account of regional variations in leukemia rates.

Whittemore confines her comments to the G and H exposure–leukemia association and discusses the evidence that this association may be causal. Specifically, she considers the evidence for each of Hill's (1965) nine criteria for causality and finds that four of the nine criteria (consistency, specificity, experiment, and analogy) are not satisfied. It is noteworthy that the conclusion that these four criteria were not satisfied could have been made without the results of our study. This illustrates that causality is very difficult to establish and that this rarely can be done on the basis of a single study. Like Whittemore, we do not regard the evidence as proving that the statistical associations we found with wells G and H were causal. What is the evidence, however, that would argue against causality? As several of the reviewers have noted, studies of occupationally exposed persons to volatile organics, TCE in particular, would not predict a detectable elevation in the risk of leukemia in persons exposed at the levels found in wells G and H in 1979. But how comforting should this be when there seems to be virtually no data on exposure to the fetus or to children and laboratory experiments show TCE to be a carcinogen and a leukemogen at high doses? The experimental evidence that would argue against volatile organics causing adverse pregnancy outcomes is even more limited, and the corresponding epidemiologic evidence for this is almost nonexistent. Thus although the available evidence does not prove causality, the statistical associations are consistent with a causal relationship and have affected our outlook on the risks to fetuses and children of exposure to these contaminants.

A practical and important question is what additional studies should be undertaken in Woburn. One suggestion we make in the article is to undertake a follow-up study of the rates of adverse health outcomes after the closure of wells G and H. Another suggestion is to replicate our study by contacting households that we did not contact. Since our survey included about a quarter of all of the pregnancies that occurred in Woburn between 1960 and 1982, it would be possible to replicate our study with one much larger. The undertaking of such an enterprise would no doubt be expensive. Because of the importance of and the concern about the health effects of contaminated water, however, the leverage that such a study could have on our views about these risks would seem to make the effort worth the expense.

Our experiences in Woburn have brought a number of methodological problems to our attention that require further work. One such area is the need for better methods for designing and analyzing studies with inexact exposure data. More accurate exposure data increase the sensitivity of statistical tests of association and results in less biased estimates of regression coefficients, but usually at an added cost that can grow very quickly. It may be possible to obtain inexpensive estimates of exposure that are reasonably efficient. For example, if the magnitude of exposure in a population can be correctly ordered (relative rank) by time and place, it may not be necessary to go to the added effort to obtain more precise exposure estimates. Determining the relative efficiencies of tests and estimates based on different exposure accuracies would help to resolve tradeoffs between added efficiency and costs. A related issue is the design of environmental health studies with inexact exposure data.

A second problem worthy of further investigation is the development of methods, similar to those in the Appendix, for the analysis of exposure–response associations in situations in which exposure can be estimated by residence history. The advantage of such methods is that they can avoid the need to sample individuals for outcome or residence information, provided that some knowledge of population densities and migration rates is available.

Finally, it would be useful to know more about the tradeoffs in selecting metrics for modeling time-dependent covariates, such as accumulating exposures. We have seen that some metrics can sometimes have very skewed distributions. In retrospect, we think a better model for our analysis of leukemia and the childhood disorders would have been one in which relative risk increases linearly with cumulative exposure, rather than exponentially. One consequence of the measure that we used is that in small samples, there can be a few very influential observations. In addition, despite the functional dependence on cumulative exposure, the resulting test could be rather inefficient for detecting a linear relationship between relative risk and cumulative exposure when the covariate distribution is highly skewed. Investigations into the relative efficiencies and robustness of metrics for time-dependent covariates with diffuse distributions would be an important contribution to the modeling literature.

ADDITIONAL REFERENCES

Christianson, R. E., van den Berg, B. J., Milkovich, L., and Oechsli, F. W. (1981), "Incidence of Congenital Anomalies Among White and Black Livebirths With Long-Term Followup," *American Journal of Public Health*, 71, 1333–1341.

Johnson, M. E., Tolley, H. D., Bryson, M. C., and Goldman, A. S. (1982), "Covariate Analysis of Survival Data: A Small-Sample Study of Cox's Model," *Biometrics*, 38, 685–698.

Myrianthropoulos, M. C., and Chung, C. S. (1974), "Congenital Malformations in Singletons: Epidemiologic Survey," in *Report From the Collaborative Perinatal Project*, ed. D. S. Bergsma, Miami: Symposia Specialists.

Warkany, J. (1971), *Congenital Malformations: Notes and Comments*, Chicago: Year Book Medical Publishers.