4

Recall Bias in Exposed Subjects Following a Toxic Exposure Incident

DANIEL G. HOPWOOD, M.P.H.
Division of Occupational and Environmental Health
San Diego University Graduate School of Public Health
San Diego, California
TEE L. GUIDOTTI, M.D., M.P.H. CCBOM
Occupational Health Program Medicine
University of Alberta Faculty of Medicine
Canada

ABSTRACT. Attempts to reconstruct health effects following toxic exposure incidents may be subject to bias from distorted or incomplete recollection. We examined recall in 22 of 31 subjects exposed to fumes from ruptured drums containing nitric acid during a U.S. Environmental Protection Agency-supervised hazardous waste site clean up operation in August 1983. Exposed subjects were interviewed by emergency room staff immediately after the episode and the next morning by telephone by a public health epidemiologist. Six months later, the subjects were again interviewed by telephone and asked to report the symptoms they had experienced during the incident. For each respondent, symptoms recalled at 6 mo were compared to symptoms reported at the time. We found a low level of agreement compared to that expected by chance, associated with significant nonrandom differences in the distribution of responses consistently favoring selective recall on the later interview. We conclude that allowing time to elapse before obtaining data on individual symptoms following exposure may lead to a significant bias in response.

THE EVALUATION of the public health risk posed by exposure to toxic substances often rests on studies of incidents involving accidental toxic exposure. Epidemiologic studies of unpredictable incidents, like all epidemiologic studies, must minimize or prevent significant bias if their results are to be accurate, precise, and generalizable. Bias is a systematic error leading to an over- or underestimation of the strength of an association between an outcome and its putative risk factors. Bias is less acceptable than random error because whereas the statistical treatment of random error underlies the techniques of biometry, bias is difficult to identify or to estimate and often consistent in leading the investigator to a false or unsupported conclusion.[1,2]

Because of the dramatic and often psychologically traumatic impact of such incidents on persons that are exposed, they would appear to be particularly subject to distortions in the recollection of the circumstances of exposure and the acute health outcomes experienced by the subjects. If recall bias is indeed a serious problem in studies of toxic exposure incidents, studies that depend on a reconstruction of the incident and its immediate consequences, particularly those of retrospective design, may be prone to limitations in accuracy and precision that would limit their usefulness.[1]

We evaluated the recall of selected symptoms among individuals involved in a two-phase toxic exposure incident in San Diego, California, in 1983 and conclude that recall bias was demonstrated in that incident. These findings raise serious questions with respect to the validity of data on health effects collected several months after an incident involving exposure to toxic substances.

Methods

During August of 1983, a hazardous materials management team under contract to the U.S. Environ-

Archives of Environmental Health
May/June 1988
V.43 no.3

mental Protection Agency was engaged in the clean up of a small abandoned dump site for chemical waste, located in southeast San Diego near the Mexican border in an industrial area distant from the nearest residential neighborhoods. Two spills occurred exactly 1 wk apart. In each case, one or more corroded 55-gallon drums of nitric acid broke open and released clouds of brownish fumes consisting of nitrogen dioxide and its colored dimer, dinitrogen tetroxide. The clouds of gas resulting were carried by low winds over several nearby auto wrecking yards, where they exposed unprotected workers, customers, and public safety personnel before the surrounding area could be evacuated.[3,4] The first spill exposed 23 individuals, including one 11-yr-old child, and the second exposed 8 additional persons. Of these, 24 were seen at one of three local hospitals; all, including those who declined medical evaluation, were interviewed the following day over the telephone by officers of the San Diego County Department of Health Services.

We obtained access to medical records from the two hospitals that received the exposed subjects and to the interview data sheets from the County telephone survey. This provided a basis for determining the initial health complaints of the exposed subjects as recorded immediately following the incident. Six months after the incident, we conducted a telephone interview with as many of the adult subjects as could be traced on three attempts. Subjects were asked to report on the symptoms they had experienced 6 mo before, at the time of the exposure incident. Because the wording and sequence of questions of the initial survey by public health officers could not be reconstructed, the wording of the later interview was based on a 20-item questionnaire based in part on a standard format developed by a committee of the American Thoracic Society for investigating respiratory disorders.[5] The survey at 6 mo contained questions regarding symptoms appropriate to nitrogen dioxide exposure[6] as well as symptoms unlikely to be associated, and included questions on occupation and whether, contrary to expectations, persistent health problems had resulted and whether medical care had been sought for these. The symptoms selected for inclusion in the survey were those itemized in the initial telephone survey for which at least two subjects indicated a positive response. If a symptom was recorded on either the emergency room treatment record or the initial survey data sheet, it was counted as having been reported at the time of the incident.

Health problems recalled in the interview at 6 mo were compared in four ways to those reported at the time of the incident.

1. The total number of paired responses to a given question which were in agreement was tallied and expressed as a proportion of all paired responses to that question. This statistic describes agreement within the group among responses to a given question in a fashion that is conceptually easy to understand but does not account for chance agreement.

2. The direction of disagreement between discrepant paired responses was examined.

3. Cohen's kappa statistic (κ) was calculated for each question, providing an index of agreement compared to that expected on the basis of chance alone. In interpreting the κ statistic, values close to 0 represent agreement no greater than that expected by chance, values between 0.2 and 0.4 suggest poor agreement, and values between 0.4 and unity suggest good agreement. Negative values suggest similar degrees of discordance.[7]

4. McNemar's χ^2 statistic was calculated for each question.[8] This statistic assumes that the reported outcomes are not independent and indicates the probability of the distribution of discordant responses occurring by chance alone. Significant bias was considered likely if $\chi^2 > 3.84$ ($p < .05$) because a discrepancy yielding a type I error of $p < .05$, if the incorrect values were accepted as true, could potentially lead to the incorrect conclusion of a statistically significant difference, but discrepancies of a lesser magnitude would not bias the ultimate conclusion to the same degree.

In interpreting the results, we considered the proportion of agreement, the direction of the discrepancy, the κ statistic as indicating net group levels of discordant responses, and the χ^2 as suggesting the magnitude of distortion of the distribution of responses within the group if the interviews were considered as replicate studies, one conducted at the time of exposure and another conducted 6 mo later.

Results

Of the 31 individuals exposed in either incident, 22 (71%) were ultimately traced and interviewed at 6 mo. The distribution of characteristics among the original and the follow-up groups is tabulated in Table 1. The age range of the original group was 11 to 52 yr with a mean age of 28.3 yr.

Of the 22 subjects interviewed at 6 mo, educational level ranged from 6 to 17 yr of education with an average of 13.2 yr. Individuals with 11 yr of education or less were generally employed in auto wrecking while paramedics (including emergency medical technicians) usually had 12 or more yr, and one had 17 yr.

The subjects who could not be traced tended to be younger and male; the average age was 23.6 yr of age, and most were employees of the wrecking yard or individuals for whom an occupation could not be determined. Of the nonrespondents, the 11-yr-old was not permitted by his parents to participate, 4 could not be reached by telephone despite an apparently valid number, 2 were known to have left the area but no number could be determined, and 2 could not be traced.

Discrepant responses as a percentage of responses to queries on all 15 symptoms ranged from 0 to 80% among the subjects. The discordance among the respondents in reporting their symptoms did not correlate with age, sex, educational level, or socioeconomic status as indicated by occupation. Certain individuals were highly consistent in their reports (7 had fewer than 20% discrepant responses), and others were highly unreliable (3 had more than 50% discrepant responses.

Table 1.—Characteristics of the Exposed Subjects

	Original group exposed		Subjects interviewed at 6 mo		Subjects lost to follow-up	
Male	26	(84%)	18	(82%)	8	(89%)
Female	5	(16%)	4	(18%)	1	(11%)
Age (mean, range)	28.3	(11–52)	31.5	(16–52)	23	(11–40)
Married	13	(42%)	12	(55%)	1	(11%)
Single	15	(48%)	8	(36%)	7	(78%)
Divorced	3	(10%)	2	(9%)	1	(11%)
White	21	(68%)	18	(82%)	3	(33%)
Black	1	(3%)	0	(0%)	1	(11%)
Hispanic	8	(26%)	3	(13%)	5	(56%)
Race-ethnicity unknown	1	(3%)	1	(5%)	0	(0%)
Wrecking yard employees	12	(39%)	9	(41%)	3	(33%)
Paramedics	10	(32%)	8	(36%)	2	(22%)
Police	4	(13%)	4	(18%)	0	(0%)
Clerical workers	2	(6%)	1	(5%)	1	(11%)
Occupation unknown	3	(10%)	—	—*	3	(33%)
Total	31	(100%)	22	(100%)	9	(100%)

Table 2.—Analysis of Recall of Reported Symptoms

	Initially reported		Recall			
Symptoms	Number	%	Proportion of agreement*	Direction†	Cohen's κ	McNemar χ^2
Dizziness	11	50	0.727	6/0	0.454	6.00‡
Headache	8	36	0.586	5/4	0.089	0.11
Lightheadedness	12	55	0.591	8/1	0.200	5.44‡
Unusual taste	12	55	0.586	9/0	0.230	9.00‡
Eye discomfort	11	50	0.632	7/1	0.276	4.49‡
Fatigue	10	45	0.727	6/0	0.421	6.00‡
Nausea	5	23	0.818	4/0	0.277	4.00‡
Pruritus	8	36	0.772	5/0	0.433	5.00‡
Abdominal discomfort	2	9	0.909	2/0	0.000‡	2.00
Paresthesias	4	18	0.773	4/1	−0.078‡	1.80
Anxiety	8	36	0.636	8/0	0.000‡	8.01‡
Respiratory						
Throat irritation	10	45	0.637	4/4	0.267	0.00
Shortness of breath	10	45	0.636	8/0	0.214	8.00‡
Cough	11	50	0.545	10/0	0.090‡	10.00‡
Sputum production	6	27	0.682	6/1	−0.085‡	4.00‡

*Proportion of subjects in which answers initially and at six months did agree.
†Pattern of discrepant responses: (− initially, + 6 mo.)/(+ initially, − 6 mo.).
‡$\kappa < 0.2$, showing little or no concordance in responses compared to chance alone. $\chi^2 \geq 3.84$ ($p < .05$), suggesting that a significant difference in results might occur had later responses been accepted as valid recall.

Table 2 presents the results of analysis of the disagreement between initial and follow-up reporting of health problems for each of 15 symptoms. Recall was subject to great variation by symptom. Only one symptom showed a high level of agreement compared to chance, without evidence for bias: headache, which is not a cardinal symptom of NO_2 exposure.[6] Even headache showed a relatively low proportion of agreement in response, however. Symptoms were reported more often in retrospect than at the time of the incident for all but one of the symptoms, throat irritation ($p < .0001$). Four (27%) of the symptoms were recalled no more often than would have been expected by chance alone. Respiratory symptoms, which would be expected as cardinal symptoms of exposure,[6] showed poorer concordance as a group than general symptoms (mean κ 0.123 vs. 0.326) and were as likely to be associated with $\chi^2 > 3.84$ (3/4 vs. 8/11).

The toxicology of the exposure incident suggests that symptoms should have abated within 1 wk.[6] Five subjects did report persistent symptoms, however. Only one reported persistence of an original symptom (shortness of breath); the others reported symptoms discrepant with their original report and likely to be associated

with unrelated personal health problems. One, who listed four persistent symptoms, was also the least precise respondent. No subject reported seeking medical care for persistent symptoms.

Discussion

We have observed substantial disagreement, exceeding that expected by chance alone and consistent in the direction of more frequent reporting of symptoms as time passed. We conclude that in this exposed population a high level of recall bias was present 6 mo following the exposure incident. Symptoms were more likely to be recalled in retrospect than forgotten. The degree of bias was similar for cardinal (respiratory) symptoms of the exposure in question and for general, often vague symptoms. The magnitude of bias is more than one might have expected given the high proportion (54%) of paramedics and police, occupations in which persons are trained to be objective in their recollection of details.

This study was able to obtain a follow-up of 71% of the exposed subjects. The subjects were not compared on a group basis, however, but to their own initial responses. The distribution of characteristics of the respondents matches those of the original group closely enough to suggest that the follow-up group was indeed representative.

The initial distribution of responses was drawn from two sources using data collected under very different circumstances, in the emergency room immediately following exposure and by telephone the day after. Because positive responses were counted if a symptom was reported in either, and negative responses were counted if denied twice or denied on the interview and not stated on the emergency room record, the combined list of symptoms initially reported is likely to be more complete and closer to the true distribution of symptoms at the time than if either source alone was used. The initial, combined list has been treated as the more valid set of data for each subject and the follow-up survey at 6 mo as an attempt to replicate the findings of the first. We have made novel use of the dependent-model chi-square statistic to suggest whether the follow-up survey would have falsely reported frequencies of symptoms significantly different from those at the time of exposure. The direction of the discrepancy in responses suggests that subjects elaborate on their experience by admitting to symptoms they had initially denied and only rarely forgot symptoms they had previously reported.

Sackett has called for empirical measurement of bias in representative study situations.[1] We echo that suggestion, particularly with respect to systematic study of the sources of bias in the investigation of accidental toxic exposure incidents. The implications of these findings are great for studies attempting to reconstruct exposure incidents. They suggest that serious levels of bias may be introduced by delay in recording symptoms and that such bias cannot be predicted either on the basis of the individual characteristics of the exposed subjects or on the basis of the expected pattern of symptoms. This is of particular importance in the reconstruction after the fact of incidents involving toxic materials or other environmental hazards. Such studies may also be prone to other sources of bias. Studies of accidents that resulted in the uncontrolled and unanticipated exposure of human beings to toxic substances are very important to our understanding of the toxicology of many agents and to risk assessment and public policy.[9] Often there are no other means by which data on human responses can be obtained. The magnitude and direction of the biases affecting these studies must be understood if they are to be designed with acceptable validity and interpreted with an understanding of their limitations.

****

We thank Drs. Craig A. Molgaard and Phillip Gay for assistance in study design, Drs. Clarke B. Hazlett and Colin Soskolne for assistance in data analysis, the San Diego County Department of Health Services, Bay General Community Hospital (San Diego), and Community Hospital of Chula Vista for providing access to medical records.

This work was supported by the Tripartite Fund for Occupational Health (University of Alberta) and by a grant from the San Diego State University Foundation.

Submitted for publication April 29, 1987; revised; accepted for publication September 29, 1987.

Requests for reprints should be sent to: Tee L. Guidotti, MD, MPH, 13-103 Clinical Sciences Building, University of Alberta Faculty of Medicine, Edmonton, Alberta (Canada) T6G 2G3.

****

References

1. Sackett, D. L. 1979. Bias in analytical research. *J Chron Dis* 32: 51-63.
2. Schlesselman, J. J. 1982. *Case Control Studies: Design, Conduct, and Analysis*, p. 179. New York: Oxford University Press.
3. Anonymous. 1983. 21 in San Diego checked after toxic gas leak. *Los Angeles Times*, 2 August, p. I-1.
4. Anonymous. 1983. Dump work sends 3 more to hospital. *Los Angeles Times*, 9 August, p. C-1.
5. Ferris, B. G. 1978. Epidemiology Standardization Project, II. Recommended respiratory disease questionnaires for use with adults and children in epidemiologic research. *Am Rev Respir Dis* 118: Suppl. 7-53.
6. Guidotti, T. L. 1978. The higher oxides of nitrogen: Inhalation toxicology. *Environ Res* 15:443-72.
7. Fleiss, J. L. 1981. *Statistical Methods for Rates and Proportions*, pp. 212-34. New York, Wiley Interscience.
8. McNemar, Q. 1947. *Psychometrika* 12:153-157.
9. Guidotti, T. L. 1984. San Diego County's Community Right-to-Know Ordinance: Case study of a local approach to hazardous substances control. *J Pub Health Policy* 5:396-409.