Correspondence

Inter-rater agreement in assessing occupational exposure in a case-control study

SIR,—I would like to draw your attention to several issues about this paper (1986;43:667-76).

(1) There are several errors in the information given in the tables. Some of these are fairly obvious—for example, in table 4 we are told that $172 \times 15 = 4500$, others were found by reworking the data.

(2) Some of the text on page 670 which describes Cohen's kappa and how it is to be interpreted is, I think, misleading.

Firstly, the authors imply that the highest possible value of Cohen's kappa is not always one; in fact it will always be equal to one when there is complete agreement between raters (when the marginal distributions must be equal). This piece of information, which is essential to any interpretation of a value of kappa, is stated in two of the references cited by the authors.^{1 2}

Secondly, the authors state that values of kappa greater than 0.5 indicate "good" to "excellent" agreement. But surely this does not imply that we should interpret values of 0.59 and 0.66, for example, as indicative of excellent agreement as the authors later claim? In the references cited by the authors to justify their interpretation it is suggested that values between 0.40 and 0.75 indicate "fair to good agreement"¹ and values between 0.40 and 0.60 "moderate"² agreement.

Ultimately, however, the seriousness of any disagreement between raters, and hence the value of indices such as Cohen's kappa, must be judged in the light of the consequences and not by any arbitrary rule. If raters disagree then at least one of them has misclassified unexposed individuals or exposed individuals, or both. The effect of such misclassification

on the estimate of the odds ratio from a case-control study is the key issue here. When the misclassification rates are the same for cases and controls, then the estimated odds ratio will be biased towards unity³—this bias can be substantial even when misclassification rates seem "reasonably" low. Without some consideration of the bias that might result from misclassification on the scale (indirectly suggested by the present study, claims that "past exposure to specific occupational agents can be assessed reasonably accurately" seem overly complacent.

ROSANNE MCNAMEE

Department of Occupational Health,
University of Manchester,
Manchester M13 9PT.

References

- 1 Fleiss JL. *Statistical methods for rates and proportions*. 2nd ed. New York: John Wiley and Sons, 1981.
- 2 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
- 3 Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol* 1977;105:488-95.
- 4 Goldberg JD. The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *Journal of the American Statistical Association* 1975;70:561-7.

Drs Goldberg, Siemiatycki, and Gérin reply:

McNamee has pointed out several inconsistencies in the tabular results that we presented in our paper. We are most grateful to her. There were errors of transcription and errors due to making up tables from computer listings generated on data sets at different degrees of completeness. None of these errors affected the message derived from the tables; the record should, nevertheless, be set straight.

Table 1—It was indicated that trial B was based on a checklist of 172 substances. In fact it was based on a checklist of 270 substances.

Tables 3, 4, 6, and 7—these should read as follows

Table 3 Pattern of agreement and selected summary statistics for trial A, in which three raters evaluated five job descriptions* using a checklist of 172 substances

No of raters attributing exposure as		No of such occurrences			
		Observed		Expected†	
Present	Absent	No	%	No	%
3	0	23	2.7	0.2	0.0
2	1	28	3.3	9.7	1.1
1	2	42	4.9	146.9	17.1
0	3	767	89.1	703.2	81.8
	Total	860‡	100	860‡	100

% Present according to§:
internal rater 1 = 9.2%
internal rater 2 = 5.8%
external rater = 4.4%
Summary statistics||:
MMAI = 95%
Kappa = 0.55 (95% CI: 0.51-0.59)

*Job descriptions were selected from those subjects in the cancer study who had been employed in the paint manufacturing industry.

†This was computed conditional on the percentage of items checked by each rater and represents the expected distribution of agreement if the exposure assessments were statistically independent.

‡860 = 5×172 .

§Proportion of exposures coded present by internal rater 1 was significantly greater ($p < 0.05$) than that coded by the other raters.

||These statistics represent crude agreement (MMAI) and chance corrected agreement (Kappa) among all three raters simultaneously.

Table 4 Pattern of agreement and selected summary statistics for trial B, in which three raters evaluated 15 job descriptions* using a checklist of 270 substances

No of raters attributing exposure as		No of such occurrences			
		Observed		Expected†	
Present	Absent	No	%	No	%
3	0	61	1.5	0.2	0.0
2	1	70	1.7	13.4	0.3
1	2	94	2.3	385.9	9.5
0	3	3825	94.4	3650.6	90.1
	Total	4050‡	100	4050‡	100
% Present according to§:		internal rater 1 = 4.0%			
		internal rater 2 = 3.6%			
		internal rater 3 = 2.6%			
Summary statistics :		MMAI = 97%			
		Kappa = 0.59 (95% CI: 0.58–0.61)			

*Job descriptions were selected from those subjects in the cancer study who had been employed in rubber products manufacturing.

†See table 3.

‡4050 = 15 × 270.

§Proportion of exposures coded present by internal rater 3 was significantly less ($p < 0.05$) than that coded by the other raters.

||See table 3.

Table 6 Pattern of agreement and selected summary statistics for trial D, in which two raters evaluated five job descriptions* using a checklist of 275 substances

No of raters attributing exposure as		No of such occurrences			
		Observed		Expected†	
Present	Absent	No	%	No	%
2	0	20	1.5	1.0	0.1
1	1	35	2.5	73.0	5.3
0	2	1320	96.0	1301.0	94.6
	Total	1375‡	100	1375‡	100
% Present according to§:		internal panel = 2.2%			
		external rater = 3.3%			
Summary statistics:		$p_0 = 97.5\%$			
		Kappa = 0.52 (95% CI: 0.38–0.66)			

*Job descriptions were generated from job function sheets supplied by an industrial hygiene department of a large chemical manufacturer.

†See table 3.

‡1375 = 15 × 275.

§Proportion of exposures coded present by the two raters differed significantly ($p < 0.05$) from each other.

Table 7 Pattern of agreement and selected summary statistics for trial E, in which two panels of raters evaluated seven job descriptions* using a checklist of 275 substances

No of raters attributing exposure as		No of such occurrences			
		Observed		Expected†	
Present	Absent	No	%	No	%
2	0	68	3.5	6.3	0.3
1	1	86	4.5	209.3	10.9
0	2	1771	92.0	1709.4	88.8
	Total	1925‡	100	1925‡	100
% Present according to§:		internal panel = 5.1%			
		external panel = 6.4%			
Summary statistics:		$p_0 = 95.5\%$			
		Kappa = 0.59 (95% CI: 0.51–0.67)			

*Job descriptions were obtained from interviews of employees currently employed at two metal fabrication plants.

†See table 3.

‡1925 = 7 × 275.

§Proportion of exposures coded present by the two raters were significantly different ($p < 0.05$) from each other.

McNamee also takes issue with some aspects of our interpretation of our results. We did not "imply that the highest possible value of Cohen's kappa is not always one." We merely pointed out that this value could not be attained when the marginal distributions are unequal. For each trial the marginal distributions as well as the kappas were presented, thereby providing complementary components of the inter-rater agreement.

The second point concerns the qualitative significance to be attached to various values of kappa. Labels such as "good," "excellent," "moderate" are arbitrary, as various authors have indicated, and depend to some extent on the subjective expectations of the investigators. Whereas we believe that in this context values of kappa in the order of 0.6 represent excellent agreement, the reader was given sufficient information to decide for herself.

The third, and most important point, concerns the connection between our results and relative risk estimates. The purpose of our case-control study was to generate hypotheses concerning past occupational exposure and cancer. There is misclassification in our exposure estimates and this misclassification leads to biased (toward the null) relative risks. It is not as important to estimate the true relative risk correctly as to determine whether or not there is excess risk. The key issue is whether the degree of misclassification is likely to be so great as to eliminate the chance of detecting excess risk. Statistical power for detecting excess risk is a function of several parameters, including the true relative risk, the degree of misclassification, and sample size. It would therefore be interesting to estimate the loss of power corresponding to levels of inter-rater agreement observed in our trials. This was beyond the scope of the paper, however. Failing such an analysis, we relied on a heuristic approach to interpreting the results. We were encouraged by finding Kappas in the order of 0.6. Furthermore, for various reasons that were outlined in the article we believe that the results of these trials provides a lower limit to the reliability of the exposure assessments carried out over a period of years in our case-control study.

We expect that the loss of statistical power corresponding to these levels of inter-rater agreement is not so great as to render the prospect of detecting excess risk a hopeless task. This expectation is premised on the unverifiable assumption that these "relatively high levels" of inter-rater agreement reflect similar levels of validity, and that such levels of validity combined with the sample sizes and appropriate values of other parameters yield adequate power for detecting excess risk.

Finally, it is important to recognise that an evaluation of our exposure assessment strategy should not

be made in isolation of the alternative methods. Previously, monitoring studies to discover occupational carcinogens have been based on subjects' job titles which were either obtained on death certificates or by more reliable means such as interviews. Although our approach entails misclassification, we are convinced that there is much more misclassification when a subject's exposure status is inferred simply from the job title than from the chemist-rater approach we have used.¹ It is in relation to the practical alternatives that we believe our approach provides reasonably accurate data. Had the results of the trials produced very low indices of agreement, we would probably have abandoned or greatly modified the project. The fact that they were not low encouraged us to carry on. We hope investigators will develop and evaluate methods of retrospective exposure assessment that would improve the prospects for case-control studies of occupational carcinogens.

Reference

- 1 Siemiatycki J. An epidemiologic approach to discovering occupational carcinogens by obtaining better information on occupational exposures. In: Harrington JM, ed. *Recent advances in occupational health*. Vol 2. Edinburgh: Churchill Livingstone, 1984:143-57.

Pulmonary fibrosis in asbestos insulation workers with lung cancer

SIR,—Kipen and colleagues reported (1987;44:96-100) that all 138 cases of asbestos insulation workers with lung cancer studied histologically had asbestosis.

It is important to note that the histological definition of asbestosis used in this study was not one which would be generally accepted by United Kingdom pathologists. The subpleural connective tissue was considered part of the interstitium for the purposes of assessing interstitial fibrosis whereas in the United Kingdom only fibrosis in the interalveolar septa or around the respiratory bronchioles and alveolar ducts is usually considered.¹ In only 130 (94%) of the cases were one or more asbestos bodies seen. Asbestosis is never diagnosed in the absence of asbestos bodies using criteria conventional in the United Kingdom where, not infrequently, the diagnosis is rejected in favour of cryptogenic fibrosing alveolitis (idiopathic pulmonary fibrosis) on the grounds that "too few" asbestos bodies are present, even when the subject is known to have been an asbestos worker.

It appears that Kipen *et al* accepted any interstitial fibrosis as evidence of asbestosis in these workers known to have had heavy exposure to asbestos. I do not intend to suggest that the fibrosis seen in these cases was not caused by exposure to asbestos but to