British Journal of Industrial Medicine 1986;43:667–676

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

M S GOLDBERG,[1 2] J SIEMIATYCKI,[2] M GÉRIN[3]

From the Department of Epidemiology and Biostatistics,[1] McGill University, Montréal, Department of Epidemiology and Preventive Medicine,[2] Institut Armand-Frappier, Laval, and Département de médecine du travail et d'hygiène du milieu,[3] Université de Montréal, Montréal, Québec, Canada

ABSTRACT The identification of occupational carcinogens in the workplace is a major concern of epidemiologists. A novel case-control approach has been developed which includes as a key component the assessment of a subject's occupational exposure history by a two stage process. Firstly, the subject is interviewed to obtain a detailed lifetime job history. Then a team of chemists and hygienists, hired and trained to do this work on a full time basis, translates each job into a list of potential occupational exposures. The present study investigated the inter-rater agreement in this type of retrospective exposure assessment. Six trials were carried out over a four year period with different raters and different sets of job files. Some trials involved only internal raters from the "in house" group whereas others involved comparisons between the internal raters and other external raters who had expertise in certain industries. In assessing exposure as simply present or absent, two summary indices of agreement were used: per cent with perfect agreement and Cohen's kappa. In most of the trials the per cent with perfect agreement among raters ranged from 95% to 98%, with kappa ranging from 0·5 to 0·7. The kappas were slightly higher for internal-internal comparisons than for internal-external ones. These results indicate a relatively high degree of inter-rater agreement and lend credibility to the validity of this type of retrospective exposure assessment.

The discovery of occupational exposures which may be harmful to health is one of the foremost problems in occupational health. Epidemiological approaches to the problem depend on the juxtaposition of the mortality or morbidity experience of workers and their occupational exposure histories. Most epidemiological case-control or surveillance studies of occupational disease have used job or industry titles as indicators for exposure. In some instances high risk occupations have been identified using this approach. The use of job titles, however, may obscure an association if only a subset of workers with a given job title were exposed to the active agent.[1 2] Statistical power will also be reduced when occupations having a common exposure are not combined.[2] Furthermore, even if an excess risk was observed it would be difficult to pinpoint the harmful agents. Ideally, then, studies should be based on chemical and physical substances to which subjects may have been exposed and not simply job or industry titles.

Since 1979, a cancer case-control study has been underway in Montreal in which detailed job exposure histories are ascertained for cases and controls.[2 3] The key element is the use of "in house" experts in chemistry and industrial hygiene to evaluate each subject's job history in order to infer a list of past exposures. Other investigators have also turned to the "expert panel" approach in determining historic exposures of study subjects in population based case-control studies[4–6] and in cohort studies.[7–9]

The credibility of results from such studies depends in large part on the accuracy of the assessment of exposure. Little has been published on the validity of retrospective assessment of occupational exposure by expert raters, mainly because this approach is comparatively recent and because of inherent difficulties of validation. The purpose of the present study was to investigate the accuracy of attributing occupational exposure by a group of chemical coders who used detailed, lifetime job descriptions obtained from in depth interviews of subjects enrolled in the cancer case-control study mentioned above (referred to as the "cancer study"). Results will be presented from a series of six trials that investigated the extent of inter-rater agreement in attributing occupational exposure

Accepted 3 March 1986

among cancer study raters, and between these raters and other chemists and hygienists from industry.

Materials and methods

CANCER STUDY

The objective of the cancer study was to discover previously unsuspected occupational carcinogens. A brief summary is given below, but the rationale and detailed methods are described elsewhere.[2 3 10-12] Men, aged 35–70, who were diagnosed in any of the major hospitals in the greater Montreal area with cancer of any of 19 sites were eligible for entry into the study. An interviewer visited the subject and conducted a two part interview, composed of a structured and a semistructured part. The structured section requested information on important potential confounders. The semistructured section elicited detailed descriptions of the occupations that each subject had held during his working life. For each reported occupation, the interviewers were trained to obtain as much information as the subjects were able to supply on the company's activities, including the raw materials; final product; machines used; any responsibility for machine maintenance; the type of room or building in which the work was carred out; the activities of workmates; the presence of gases, fumes, or dusts; and any other information that could furnish a clue regarding possible chemical or physical exposures. Each completed questionnaire was examined by a team of coders, composed of chemists, industrial hygienists, and engineers, who then translated each reported occupation into a list of potential exposures. The team of coders relied on the following sources as a basis for estimating exposure: their own industrial experience and chemical knowledge, old and new technical and bibliographical material describing industrial processes, and consultation with experts familiar with particular industries. The coding of exposure was done on a checklist that contained chemical and physical exposures (to be referred to as "substances"). Any other substances believed to have been present could also be added to the list. The composition of the checklist has evolved over the period of the cancer study. Initially, there were 172 substances; then for a period there were 259; a year later there were 270; and now there are 275. In the day to day coding (referred to as the "routine coding") exposure was attributed through a type of consensus process whereby one coder intially attributed exposure and one or more other chemical coders reviewed the original codings.

For each job description in each study subject's work history, the project chemists indicated whether any of the substances on the checklist may have been present in the worker's environment. For each substance which was checked off as possibly present, the coders further indicated: (*a*) their confidence that the exposure had actually occurred ("possible," "probable," or "definite"); (*b*) the average level of concentration ("low", "medium," or "high"); (*c*) the frequency of exposure during a normal workweek ("less than 5%," "5–30%," or "more than 30%"), and (*d*) the type of contact ("respiratory," "cutaneous," or "both").

INTER-RATER AGREEMENT TRIALS

True validations were impossible because of the absence of valid historical information on the exposures of the study population. The trials of inter-rater agreement were intended, nevertheless, to shed light on the accuracy of the chemical coder's exposure assessments. Some of the trials were designed to determine the degree of agreement in coding among the chemical coders of the cancer study (hereafter referred to as "internal raters"), whereas others were meant to compare the assessments of the internal raters with those of external raters. The latter were

Table 1 *General characteristics of the agreement trials*

		Job descriptions examined				*No of raters*	
Trial	*Date*	*Industry/occupation*	*Selected or obtained from*	*No*	*No of substances on exposure checklist at that time*	*Internal*	*External*
A	1981	Paint manufacturing	Cancer study	5	172	2	1
B	1981	Manufacturing of rubber products	Cancer study	15	172	3	0
C	1982	Welding and soldering	Cancer study	18	259	2	1*
D	1984	Chemical manufacturing	Company records	5	275	1†	1
E	1984	Manufacturing of metal products	Interviews	7	275	1†	1
F	1984	Miscellaneous assortment	Cancer study	23	270‡	2%	0

*Three external raters coded exposure as a group.
†A panel of two internal raters attributed exposure as part of the routine coding.
‡Number of substances on the checklist at the time of the first coding.
%Two routine codings by two teams of internal raters were performed at two different times.

experts in various industrial sectors, engaged solely for the purpose of participating in one or another of the trials. In total, six trials of inter-rater agreement were carried out between 1980 and 1984; the main features of each of the trials are presented in table 1, in chronological order of the completion of the trials.

Owing to personnel turnover, different internal raters participated in the various trials. These raters had different professional backgrounds and varying degrees of coding experience. As indicated above, the coding checklist evolved over time and the various trials involved versions containing 172, 259, 270, and 275 substances.

Since the cancer study is a population based case-control study, the whole range of occupations and industries in the Montreal area appears among study subjects. For most of the trials the job descriptions that were evaluated were taken from among cancer study subjects. For two of the trials (D and E), job files were generated from outside sources—one from hygiene records of a Canadian chemical manufacturer and one from interviews done with metal workers whose working environment had recently been measured by hygienists.

For each trial, a set of job descriptions was selected according to the criteria outlined below, and the various raters participating in the trial independently coded exposure for each job using the exposure checklist in use at the time of the trial. In some instances one of the "raters" was a team operating by consensus. Whenever external raters participated they were trained in the methodology and criteria of the coding process, though care was taken not to influence their substantive judgements. More details about each trial are presented below.

STATISTICAL METHODS

The primary goal of the analyses was to determine the extent of agreement between raters in identifying the presence or absence of exposure. Thus for each trial agreement was assessed on a binary scale (present, absent). The methods used in these analyses, however, were extended, as described below, when, for example, agreement was assessed for concentration of exposure on a four-point measurement scale.

Suppose that n job descriptions were examined by r raters, coding exposures on a checklist of s substances. For each of the s substances it was theoretically possible to assess the agreement among r raters on the n job descriptions. Determining agreement on a substance by substance basis posed two problems: (*a*) since the number of job descriptions, n, was small (from 5 to 23, depending on the trial), the variability of any agreement statistic concerning a specific substance was quite large and (*b*) since the number of substances, s, was large (from 172 to 275, depending on the trial), it would be difficult to assimilate the multitude of results. Thus some method of summarising agreement over the s substances had to be found. To this end, the n × s decisions of each rater were treated as independent observations and the agreement in rating among r raters was assessed over these n × s observations.

For each trial a 2^r table—that is, 2 × 2 for a two rater trial, 2 × 2 × 2 for a three rater trial—was derived. Each of the n × s entries were classified according to the combination of raters coding them present or absent. For instance, the data layout for trial A in which there were three independent raters is presented in table 2. There were 860 (5 × 172) observations distributed among the eight combinations in this table.

The data in the 2^r table was first synthesised by counting the number of distinct patterns of agreement, ignoring the identities of the raters. Based on the marginal proportions for each rater, an expected number for each pattern of agreement was estimated assuming the null hypothesis that the raters' codings were statistically independent of each other. Table 3 shows the results of such an analysis for trial A, as derived from table 2. Such a table provides a fairly complete, yet concise, description of the results.

Several parameters were then estimated from this summary table and, for this example, from the three 2 × 2 tables derived from pairwise comparisons. For each rater, the number of substances attributed was summed over all n job descriptions and then divided by the total possible number of such assignments (n × s). (This proportion will be referred to as the "proportion of exposures attributed.") Equality of the three proportions was tested using Cochran's Q-index.[13] For each pair of raters, the equality of the two proportions was tested using McNemar's test.[14] Both of these statistics are asymptotically dis-

Table 2 *Detailed pattern of agreement in trial A, in which three raters examined five job descriptions using a checklist containing 172 substances*

Substance rated as:			
Rater 1 (Internal)	*Rater 2 (Internal)*	*Rater 3 (External)*	*No of occurrences*
p	p	p	23
p	p	a	19
p	a	p	9
a	p	p	0
p	a	a	28
a	p	a	8
a	a	p	6
a	a	a	767
			Total 860*

p = Exposure was "present."
a = Exposure was "absent."
*860 = 5 × 172 = number of observations.

tributed as chi-squared, with the number of degrees of freedom equal to the number of raters less one.

Among several measures of agreement available for the simple two dimensional tables,[13,15] two were chosen for presentation: the index of crude agreement,[16] p_0, which simply equals the proportion of observations for which all raters agreed that exposure was present or absent, and Cohen's kappa index.[15,17] The latter explicitly accounts for the extent of agreement expected on the basis of chance, p_e, assuming statistical independence of the raters' codings. Kappa is calculated by subtracting expected from crude agreement ($p_0 - p_e$) and then dividing by the maximum expected excess agreement ($1 - p_e$). Thus a value of kappa of zero indicates a degree of agreement that would be expected by chance ($p_0 = p_e$). The maximum value of kappa is dependent on the marginal distributions. When these are equal, kappa can attain a maximum value of unity; otherwise, it is less than one.[17] Values of kappa greater than 0·5 are considered to represent "good to excellent" agreement.[15,18] For trials in which the coding of more than two raters was compared simultaneously, a generalised version of the index of crude agreement, known as the mean majority agreement index (MMAI),[19] was used. This index is equal to an average, over the n × s observations, of the proportion of ratings in which there was unanimous agreement for each exposure assignment. It has a maximum value of 100%. Furthermore, a generalised version of kappa,[15] based on the intra class correlation coefficient[16] was also used. The pairwise agreement statistics and related 95% confidence intervals (95% CI) were computed by means of a FORTRAN program (M S Goldberg, unpublished observations).

Detailed results will be presented for each trial based on classifying exposure as either "present" or as "absent." Thus a substance labelled as present in these analyses could have been coded with any level of dose, as defined by the various combinations of frequency and concentration. Other analyses were in fact carried out but will be presented in summary form only. One set was similar to the present/absent analysis but used a different threshold for distinguishing present from absent—namely, only those exposures which were rated as "definitely present" were counted as present. Another set of analyses concerned the extent of agreement in rating concentration, frequency, and type of contact.

Description and results of each trial

TRIAL A

Trial A was a comparison between two cancer study raters and an outside expert. The external rater was an industrial chemist who had more than 10 years experience in the paint manufacturing industry. Unlike the routine coding procedure, the internal raters coded these files independently without a consensus rating. Thus the statistical comparisons were between the two internal raters and between each of them and the external rater. Five job descriptions of paint industry workers were selected from the bank of cancer study files.

The results from this trial are shown in tables 2 and 3. Internal rater 1 attributed a significantly larger number of exposures than either of the other two raters ($p < 0·05$). Nevertheless, the observed pattern of agreement was much better than that expected by chance alone. A kappa of 0·66 and an MMAI of 95%,

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	0·1	0·0
2	1	28	3·3	5·6	0·7
1	2	42	4·9	125·6	14·6
0	3	767	89·1	728·7	84·7
	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 = 0·66 (95% CI: 0·62–0·69)			

*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 172 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	13·8	0·3
1	2	94	2·3	393·0	9·7
0	3	3825	94·4	3643·0	90·0
	Total	4500‡	100	4500‡	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.
‡4500 = 15 × 172.
§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.

both indicative of excellent agreement, were observed.

TRIAL B

Trial B concerned concurrent independent coding by three of the cancer study raters. Fifteen job descriptions were selected from the cancer study, all of rubber industry workers.

The results of this trial are presented in table 4. Two of the raters attributed about the same number of exposures; the third, however, coded fewer. As in the previous trial, the number of observed exposures in which there was perfect agreement was greater than that expected by chance. This was reflected in the summary indices which indicated excellent agreement (MMAI = 97%; kappa = 0·59).

TRIAL C

In trial C there were three external raters who attributed exposure as a team. This group of raters consisted of an engineer engaged in welding research at a Montreal university, a welder employed in a trade association, and a welding instructor with over 30 years experience in industry. Two internal raters attributed exposure independently of each other. For this trial 18 job descriptions of welding and soldering workers were selected from the files of the cancer study.

The results for this trial are shown in table 5. There were significant differences among raters in the proportion of exposures attributed. Nevertheless, the extent of crude agreement (MMAI = 93%) and

Table 5 *Pattern of agreement and selected summary statistics for trial C, in which three raters evaluated 18 job descriptions* using a checklist of 259 substances*

No of raters attributing exposure as		*No of such occurrences*			
		Observed		*Expected†*	
Present	*Absent*	*No*	*%*	*No*	*%*
3	0	196	4·2	2·5	0·1
2	1	162	3·5	85·4	1·8
1	2	237	5·1	968·6	20·8
0	3	4067	87·2	3605·5	77·3
	Total	4662‡	100	4662‡	100

% Present according to§: internal rater 1 = 9·7%; internal rater 2 = 8·4%; external panel = 6·6%

Summary statistics‖: MMAI = 93%; Kappa = 0·62 (95% CI: 0·60–0·64)

*Job descriptions were selected from those welders who had been interviewed during the course of the cancer study.
†See table 3.
‡4662 = 18 × 259.
§Proportion of exposures coded present differed significantly ($p < 0·05$) among all three 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	0·8	0·0
1	1	35	2·5	67·1	4·9
0	2	1320	96·0	1307·1	95·1
	Total	1375‡	100	1375‡	100
% Present according to§:		internal panel = 2·0% external rater = 3·0%			
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 = 5 × 275.
§Proportion of exposures coded present by the two raters differed significantly ($p < 0·05$) from each other.

chance corrected agreement (kappa = 0·62) were high.

TRIAL D

Trial D was carried out with the collaboration of the industrial hygiene department of a large Canadian chemical manufacturer. They selected for this trial a plant located in another Canadian city, in which detailed job function and hygiene descriptions had been recorded for their own internal use. Interviewing workers was impossible because of the distance. Instead, job descriptions were generated, using the information in the job function sheet and a company hygienist's personal knowledge of the workplace. One of the cancer study interviewers, the hygienist and one of the authors (MSG) created job descriptions that workers might have been expected to provide through interviews. Five such job descriptions were developed, based on jobs in different parts of the plant. These were put into the routine coding work of the team of internal raters and they carried out their usual consensus rating procedure, unaware that these jobs were not really part of the cancer study. The company hygienist was the external rater.

There were significant differences in the two estimates of the proportion of exposures attributed present ($p < 0·05$), with the internal panel coding one third as many exposures as the external rater (table 6). Crude agreement was high (p_0 = 97·5%), however, and was significantly greater than that expected by chance (kappa = 0·52).

TRIAL E

Trial E exploited the fact that hygienists working in a

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	56	2·9	5·3	0·3
1	1	84	4·4	191·6	9·9
0	2	1785	92·7	1728·1	89·8
	Total	1925‡	100	1925‡	100
% Present according to§:		internal panel = 4·7% external panel = 5·8%			
Summary statistics:		p_0 = 95·6% 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.

local community health department had recently conducted hygiene surveys in selected local industries. Among these industries were two metal fabrication plants. As a result of this programme, detailed measurements of concentrations of some airborne materials in the vicinity of each worker were obtained. The survey also provided information on materials and equipment used by each employee. A chemical breakdown of some of the materials used was also obtained from some manufacturers.

Two of the hygienists who had conducted these surveys agreed to participate in the trial. Seven current employees at the two plants were interviewed by one of the cancer study interviewers to obtain a current job description—similar to those routinely obtained in the course of the cancer study. One of the subjects was a drill press operator, one a metal polisher, one a painter, two were welders, and two metal forming workers. The resulting job descriptions were coded by the two hygienists who had carried out the detailed surveys and their evaluations provided one consensus rating. As in the previous trial, the job descriptions were placed in the routine coding of the team of internal raters.

Table 7 shows the detailed and synthesised results. Although there were significant differences in estimating the total prevalence of exposure ($p < 0.05$), there was considerable overlap in the exposures attributed by the two raters ($p_0 = 95.6\%$; kappa = 0.59).

TRIAL F

Trial F was an outgrowth of trials D and E. Normally, in the cancer study, the interview provided a job description for each job that the subject had had in his working life. For trials D and E, the external raters had expert knowledge only of the worker's current job, and only these were obtained in conjunction with the external raters. Had only the current job description for a given worker been provided, the team of internal raters would have realised that a special investigation was under way. To mask this, a complete lifetime job history was generated for the 12 job descriptions that had formed the basis of trials D and E. This was done by adding to the current job description one or two others which were selected from among the thousands of job descriptions on hand in the cancer study—that is, one or two job descriptions already coded by our internal raters at some time in the preceding five years were copied out by an interviewer as if they were part of a recent interview. The internal raters did not recall that these job descriptions had already appeared in the study. A total of 23 job descriptions that had been rated previously by the internal raters were recoded in trials D and E. It was not necessarily the same individuals who rated the 23 jobs on the two separate occasions. This was because of changes in personnel through the course of the cancer study and because different subsets of raters would assess different files in the routine coding. Nevertheless, the trial represents a comparison of the routine coding at two points in time, thus providing what might be termed a test of reliability of the coding system. In addition, because of changes in the number of substances under consideration, the first coding of these files used an exposure checklist containing 270 substances whereas the second was based on a list of 275 substances. Thus the trial was analysed using 270 substances.

As shown in table 8, the two panels of internal raters coded nearly identical numbers of exposures. High levels of crude ($p_0 = 98.5\%$) and chance corrected agreement (kappa = 0.67) were observed.

Table 8 *Pattern of agreement and selected summary statistics for trial F, in which two panels of internal raters evaluated 23 job descriptions* using a checklist of 270 substances*

No of raters attributing exposure as		*No of such occurrences*			
		Observed		*Expected†*	
Present	*Absent*	*No*	*%*	*No*	*%*
2	0	99	1·6	3·4	0·0
1	1	94	1·5	285·2	4·6
0	2	6017	96·9	5921·4	95·4
	Total	6210‡	100	6210‡	100

% Present according to: internal panel 1 = 2·3%; internal panel 2 = 2·4%

Summary statistics: $p_0 = 98.5\%$; Kappa = 0·67 (95% CI: 0·61–0·73)

*Job descriptions were selected from the entire pool of files on hand in the Cancer Study that had been previously attributed in the routine coding.
†See table 3.
‡6210 = 23 × 270.

Table 9 *Selected summary statistics of agreement for all pairwise comparisons in all trials of inter-rater agreement*

		Statistics of agreement for					
		Internal-internal pairs			*Internal-external pairs*		
Trial	*No of observations**	*No of pairs of of raters*	*Average p_0 %*	*Average Kappa*	*No of pairs of of raters*	*Average p_0 %*	*Average Kappa*
A	860	1	97·8	0·62	2	95·5	0·51
B	4050	3	97·3	0·59	0	—	—
C	4662	1	97·9	0·69	2	95·5	0·58
D	1375	0	—	—	1	97·7	0·52
E	1925	0	—	—	1	95·9	0·59
F	6210	1	98·5	0·67	0	—	—
	Average or total:	6	97·8	0·63	6	95·9	0·55

*Equal to the number of job descriptions multiplied by the number of substances on the relevant coding checklist.

ADDITIONAL RESULTS

Some trials compared internal raters with each other, some compared internal with external raters, and some did both. Table 9 shows, in synthesised format, the summary agreement indices between all pairs of raters, distinguishing the internal-internal comparisons from the internal-external comparisons. Crude and chance corrected agreement was somewhat higher for comparisons among internal raters than for internal-external ones.

In five of the trials there were statistically significant differences between raters in the proportion of exposures attributed. When the same analyses were run, however, using a more stringent threshold for classifying exposure as present—namely, definitely present—then differences in the proportion of exposures attributed were considerably reduced. None the less, using this more stringent threshold, values of p_0 and kappa were virtually unchanged from those presented in the present/absent analyses.

For each substance checked off as present, the rater had to indicate on three point scales the concentration and frequence of exposure and whether the exposure was cutaneous, respiratory, or both. To assess agreement, absence of exposure was added as another category to these scales, yielding, in effect, three four-category variables: concentration, frequency, and type of contact. Analyses on these four point scales for each of these three variables were conducted, analogously to those carried out for the binary present/absent variable. Averaged over all pairwise comparisons in all trials, the following results were obtained: concentration, average kappa = 0·41; frequency, average kappa = 0·46; and type of contact, average kappa = 0·52. As in the absent/present analysis, agreement among internal raters was slightly higher than that observed between internal and external raters.

Discussion

Inter-rater agreement was assessed by three general approaches: (1) comparison of the different raters' estimates of proportion of exposures attributed (using McNemar's test and Cochran's Q-index), (2) comparison of observed and expected distributions of rater agreement, and (3) estimation of crude and chance corrected agreement statistics (p_0 and kappa). The interpretation of the significant differences in proportions of exposures attributed is difficult because the tests used are sensitive and might detect real but small effects. Indeed, although the McNemar tests usually indicated significant differences in exposure prevalence—that is, proportions of substances attributed—the differences were generally small. In only one case (see table 3) was there a serious discrepency between the raters. We believe that in evaluating inter-rater agreement, emphasis should be placed on the crude and chance corrected indices—that is, p_0 and kappa—and on the comparison of observed and expected patterns of agreement.

Despite differences in estimates of exposure prevalence between some raters, both crude and chance corrected statistics indicated, in all trials, that there was "good" to "excellent" agreement in identifying exposures. Furthermore, good agreement was observed in coding frequency, concentration, and type of contact of exposure. The kappa statistics from these latter analyses indicated somewhat lower agreement than that observed in the present/absent analyses. This was expected since these exposure indices were evaluated on four point scales, thus allowing a greater chance for disagreement. Over all, the results were encouraging.

One theoretical limitation to the use of these methods is that expected values were calculated under the assumption that each rater's various assessments

are independent of each other. Strictly speaking, this assumption was violated, because a rater's coding of one substance in a job description may well have influenced his judgement on others. It is unlikely, however, that the results were appreciably affected.

These trials were intended to investigate the validity and reliability of retrospectively assessing occupational exposure by technically trained personnel. The evidence regarding validity is indirect since none of the trials was based on a "gold standard" by which to judge the coding of the cancer study raters. In some trials the external raters' coding may be considered to be closer to the truth than that of the internal raters, but in no instance can there be assurance that any of the external raters had perfect knowledge of exposure. Indeed, it should not be assumed that the external raters necessarily provided more valid codings than the internal raters. While the external raters were expert in specific industrial processes, they were not experienced in evaluating workers' exposures. Thus the lower degree of agreement between internal and external raters as compared with that among internal raters does not necessarily indicate that the coding of the internal raters was less valid. The higher agreement among internal raters, however, may be partly explained by the fact that, as a result of their close working arrangement, common attitudes were engendered and a common base of knowledge was developed over time. This is best exemplified by the results of trial F, in which a high degree of agreement was observed between two panels of internal raters who, using the cancer study consensus approach, attributed exposure to the same set of files on two different occasions.

The four trials comparing the coding of internal and external raters showed that, for both individual and teams of coders, there was good agreement over a range of industries and occupations and time periods of employment. In trial E routine cancer study coding was compared with that of two industrial hygienists for a set of current occupations in the metal fabrication industry. The exposure data used by the external raters was obtained from a fairly thorough occupational survey, whereas in the other trials reliance was placed on the raters' personal knowledge. In trials A and C assessments of exposure were made by individual raters for paint manufacturing and welding occupations in which employment had occurred in the past. In trial D job descriptions were generated from the hygiene records of a chemical manufacturer. The job descriptions were fictitious and thus were of poorer quality than those of the other trials. Furthermore, the type of industrial operation of the manufacturer is not found in the cancer study's catchment area, so that the internal raters had virtually no experience with the specific chemical processes used. Nevertheless, a relatively high degree of agreement between the internal and external raters was observed.

The industries chosen as a focus for the trials—paint, chemical, metal, welding—are among the most complex and difficult for the chemical coding of exposure. A general population case-control study, such as the cancer study, would include a cross section of jobs and industries that on average would be easier to code than those evaluated in most of the trials of this report. Thus the levels of inter-rater agreement shown here probably represent lower limits of inter-rater agreement for the whole range of dossiers in the cancer study. Combining the results of this study with the previous finding that there is no substantial underreporting of jobs by interviewees,[20] it appears that the translation into exposures of a worker's reported job history by a group of chemists and hygienists provides reasonably valid data.

There have been few other reports of validations of retrospective occupational exposure assessments made by expert raters. In a study of petrochemical workers[8] exposure to eight substances was assessed by company engineers from a dictionary of job titles which contained the date and area in the plant of employment for each employee. Much lower levels of agreement were observed there than in the present study. This may be attributed to the lack of detail in the job title and, perhaps, to an inconsistent use of the coding criteria. In a prospective monitoring programme of a chemical plant,[7] exposure coded by company engineers was indirectly validated by verifying that those workers diagnosed with angiosarcoma were attributed higher levels of exposure to vinyl chloride, a known liver carcinogen, than a group of controls.

The results from this study lend credibility to the notion that past exposure to specific occupational agents can be assessed reasonably accurately. Several questions remain outstanding. For example, it is not clear whether validity is affected by the nature of the job descriptions available to the raters and by the skill and experience of the raters. Also of importance is the degree to which agreement may vary from substance to substance. More work is clearly required to assess this potentially valuable epidemiological instrument.

We are grateful to Drs Graham Gibbs and Gilles Lebeau and Mr Jacques Guénette for permitting access to industrial sources of data. Chemical coding in the cancer study was performed by Denis Bégin, Dr Joseph Hubert, Howard Kemper, Lucien Laroche, and Christian Millet. Many thanks to Jean-Paul Boillot, J Doré, Michel Galopin, Denis Lessard, Peter Nichol, Roger Tremblay, and Peter Wrzesien for their time and effort in performing the external codings.

Denise Bourbonnais performed the interviews and selected job descriptions that were used in three of the trials. Much of the data management was supervised by Lesley Richardson who has also contributed to various aspects of design. We appreciate the helpful comments of Nancy Mayo and many thanks to Claudette Richer for preparing the manuscript.

This research was supported by grants from the Institut de recherche en santé et en sécurité du travail du Québec, the National Health Research and Development Program, and the National Cancer Institute of Canada. Mr Goldberg gratefully acknowledges receipt of personal support from the Institut de recherche en santé et en sécurité du travail du Québec, the Fonds FCAC pour l'aide et le soutien à la recherche, and McGill University.

References

1 Hoar SK, Morrison AS, Cole P, Silverman DT. An occupational and exposure linkage system for the study of occupational carcinogenesis. *J Occup Med* 1980;**22**:722–6.

2 Siemiatycki J, Day NE, Fabry J, Cooper JA. Discovering carcinogens in the occupational environment: a novel approach. *JNCI* 1981;**66**:217–25.

3 Siemiatycki J. An epidemiological approach to discovering occupational carcinogens by obtaining better information on occupational exposures. *Recent Advances in Occupational Health* 1984;**2**:143–57.

4 Macaluso M, Vineis P, Continenza D, Ferrario F, Pisani P, Andisio R. Job exposure matrices: experience in Italy. In: Acheson ED, ed. *Job exposure matrices*. Southampton: Medical Research Council, 1982:22–30.

5 Hernberg S, Collan Y, Degerth R, *et al.* Nasal cancer and occupational exposures: preliminary report of a joint Nordic case-referent study. *Scand J Work Environ Health* 1983;**9**:208–13.

6 Hernberg S, Korkala N, Asikainen U, *et al.* Primary liver cancer and exposure to solvents. *Int Arch Occup Environ Health* 1984;**54**:147–53.

7 Greenberg RA, Tamburro CH. Exposure indices for epidemiologic surveillance of carcinogenic agents in an industrial chemical environment. *J Occup Med* 1981;**23**:353–8.

8 Soskolne C. *Upper respiratory cancer among refinery and chemical plant workers: a case-control study in Baton Rouge, Louisiana.* Philadelphia: University of Pennsylvania, 1982. (PhD dissertation.)

9 Bond G, Sobel W, Shellenberger R, Flores G. Validation of work histories obtained from interviews. *Am J Epidemiol* 1985;**122**:536–7.

10 Siemiatycki J, Gérin M, Hubert J. Exposure-based case-control approach to discovering occupational carcinogens: preliminary findings. In: Peto R, Schneiderman M, eds. *Quantification of occupational cancer.* (Banbury report No 9.) New York: Cold Spring Harbor Laboratory, 1981:471–83.

11 Siemiatycki J, Gérin M, Richardson L, Hubert J, Kemper H. Preliminary report of an exposure-based, case-control monitoring system for discovering occupational carcinogens. *Teratogenesis Carcinog Mutagen* 1982;**2**:169–77.

12 Gérin M, Siemiatycki J, Kemper H, Bégin D. Obtaining occupational exposure histories in epidemiologic case-control studies. *J Occup Med* 1985;**27**:420–6.

13 Landis JR, Koch GG. A review of statistical methods in the analysis of data arising from observer reliability studies, parts 1 and 2. *Statistica Neerlandica* 1975;**29**:101–23, 151–61.

14 Bennett BM. Tests of hypotheses concerning matched samples. *Journal of the Royal Statistical Society B* 1967;**29**:468–74.

15 Fleiss JL. *Statistical methods for rates and proportions.* 2nd ed. New York, John Wiley and Sons, 1981.

16 Rogot E, Goldberg ID. A proposed index of measuring agreement in test-retest studies. *J Chronic Dis* 1966; **19**:991–1006.

17 Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960;**20**:37–46.

18 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.

19 Armitage P, Blendis LM, Smyllie HC. The measurement of observer disagreement in the recording of signs. *Journal of the Royal Statistical Society A* 1966;**129**:98–109.

20 Baumgarten M, Siemiatycki J, Gibbs GW. Validity of work histories obtained by interview for epidemiologic purposes. *Am J Epidemiol* 1983;**118**:583–91.