Self-assessed Versus Expert-assessed Occupational Exposures

Lin Fritschi,¹ Jack Siemiatycki,¹ and Lesley Richardson¹

While self-response to a checklist of substances may be a convenient and inexpensive method for obtaining information on occupational exposure, the validity of such information has not been evaluated. The objective of this report is to provide some evidence concerning validity of self-reported occupational exposures. In the context of a large case-control study, it was possible to compare self-reports with expert assessment, in which a team of industrial hygienists and chemists examined each job history individually and decided on likelihood of exposure. The subjects were 1,910 males who had participated in a population-based case-control study of cancer and occupational exposures conducted in Montreal, Canada, between 1979 and 1985. For each of 11 substances, the two methods of exposure assessment were compared by means of a kappa statistic and by computing the sensitivity and specificity of self-assessment against expert assessment. Kappa values ranged from 0.33 to 0.64. Compared with the expert assessment, specificities of the self-assessment were generally high (0.83–0.97, with a median of 0.90), but sensitivities were low (0.39–0.91, with a median of 0.61). The authors conclude that self-reports of occupational exposure are not sufficiently accurate to warrant their sole use in most community-based studies. Am J Epidemiol 1996;144:521–7.

epidemiologic methods; occupational exposure

Occupational risk factors for cancer can be investigated in industry-based cohort studies or community-based case-control studies. Each approach has advantages and disadvantages. A cohort study may have access to exposure records, but the small numbers of cases often limit the usefulness of this approach. Case-control studies can include much larger numbers of cases, but typically attempt to obtain exposure information retrospectively without the benefit of reliable records. In the past, case-control study subjects were asked to provide job histories, and the job titles served as the exposure variables for epidemiologic analyses. With the growing recognition that job titles may be poor proxies for occupational exposures (1, 2), there have been attempts to use different methods to obtain information regarding past exposure to specific occupational substances. At one extreme are studies in which study subjects are asked whether they have been exposed to particular substances, and their responses are used as the variables for analysis. This is a relatively simple and inexpensive way of collecting occupational data. Another approach that has been developed is the job-exposure matrix, which is a database that can be used to infer exposures from a job title. This approach can also be relatively cheap, but it depends on the availability of a relevant and valid job-exposure matrix and also does not address the issue of variability of exposures within a job title. In a third method, experts in exposure assessment evaluate job histories provided by study subjects and attribute exposures to each subject individually. The decision about which method to use is determined by many factors. If the least expensive method, self-assessment, can be shown to provide reasonably valid data, then this would be an attractive option for many investigators.

This report describes a comparison of two methods of assessing occupational exposure to 11 substances: self-assessment by written questionnaire and expert-assessment, in which a team of industrial hygienists and chemists examined each job history individually and decided on likelihood of exposure.

MATERIALS AND METHODS

A case-control, population-based study was conducted in Montreal, Canada, to look for relations between 19 cancer sites and 294 occupational exposures. The study has been described in detail elsewhere (3), and only a brief description will be given here. Cases were all incident cases of cancer that occurred in male
Montreal residents between 1979 and 1985. Some population controls were selected by random digit dialing, and some were randomly selected from electoral lists.

Subjects were sent a brief, self-administered questionnaire in which they were asked to summarize their employment history. They were also asked whether they had ever been exposed to any of 13 different groups of substances. The wording of the question was, "Have you ever worked in the production of the following substances, or have you ever used any of them in your work (taking into account all of your previous jobs)?" For each question, there was a "Yes" or a "No" box to be ticked. The purpose of this self-administered questionnaire was to "prime the pump" for the interview that was to follow in a few days. It was to get the subject thinking about his job history and the exposures he might have experienced. Rather than specific chemicals, we chose broad use categories such as solvents, pesticides, and wood. A few days after delivery of this questionnaire to the subject, an interviewer was dispatched to visit and carry out an interview. If the self-administered questionnaire had been completed, the interviewer collected it and used the employment history as a basis for a probing, in-depth questionnaire regarding the industrial process and working conditions for each job in the man's history. If the questionnaire had not been completed, the interviewer elicited the employment history and then carried out the in-depth interview, but did not bother with the checklist of substances. The subject was also asked his age, ethnicity, income, and smoking history.

After the interview, a team of chemists and hygienists examined each completed questionnaire and translated each job into a list of potential substance exposures by using a checklist of 294 substances (3, 4). For each substance, the experts noted their degree of confidence that the exposure had actually occurred (low, medium, or high probability of exposure). The experts based their assessments on the job descriptions elicited from the respondents, including their self-reported exposures, scientific and technical literature on industrial processes and industrial hygiene, consultations with selected local experts in particular industrial sectors, and their own judgment and knowledge. Although the self-reported exposures constituted one of the inputs to the assessment, our experts only coded an exposure if they had reason to do so independently of the respondent's self-report.

The present methodological analysis is based on the comparisons between the self-assessed exposure to the substances and the expert-assessed exposure to the approximately equivalent substances. For comparison of the two types of assessment, it was necessary to make adjustments in three respects. First, for some of the 13 substances on the self-report checklist, there was no exact match between the informal name for the substance used in the self-administered questionnaire and the more technical terms used on the checklist of 294 substances coded by the experts. Consequently, we tried to find the group of substances on the detailed checklist that best corresponded to the category on the self-report checklist. For two categories on the self-report checklist, there was no satisfactory match, and these were dropped. Table 1 shows the 11 remaining terms on the self-report checklist and their corresponding exposure variables from the detailed checklist.

Second, the self-report was based on the response to a single question covering the entire working history, whereas the experts assessed exposures on a job-by-job basis. Thus, for this analysis, the expert assessments were collapsed so that they corresponded to the "ever exposed/never exposed" quality of the self-reports.

Third, whereas the self-report was a simple dichotomy, the experts used a multidimensional, semiquantitative rating system. For our purpose, we collapsed the expert rating into a dichotomy analogous to the self-report. Any level of exposure was retained as exposed except if the expert coded the level as low probability of exposure. In that case, we did not count the subject as exposed or unexposed; rather, we excluded him from the analysis.

Two types of indices of agreement were computed. The first type, kappa, makes no hierarchic assumptions between the two ratings. The second type was based on the premise that among the two ratings the expert-assessed exposure was closer to the truth. On this basis, we computed the sensitivity and specificity for the self-assessed exposure. We repeated the analyses separately for the population controls and the cancer cases. Analysis was performed using SPSS (5).

RESULTS

There were 5,316 subjects eligible for the cancer study. Of these, 4,263 were interviewed, but only 1,910 (45 percent) completed the self-assessed exposure questions. Compared with the rest, those who completed the self-assessed exposure questions were significantly (p < 0.01) more likely to have higher incomes and to be French-speaking, were less likely to have had a proxy to respond, and smoked, on average, more cigarettes over their lifetime. Of the 1,910 subjects who completed the self-assessed exposure questions, 1,657 were cancer patients and 253 were population controls.

Am J Epidemiol Vol. 144, No. 5, 1996
For each substance, self-assessed and expert-assessed exposures were compared for all the subjects who completed the self-assessed exposure questions after the exclusion of the few subjects whom the experts coded as having low probability of exposure. Kappa values for these comparisons ranged from 0.33 for plastic or rubber to 0.64 for fur or leather (table 2). When the expert assessment was assumed to represent the "gold standard," the specificities of the self-assessed evaluation were over 0.80 for all of the substances (median, 0.90). However, the sensitivities were lower and more variable (median, 0.61).

The analyses were repeated after excluding those cases that the experts coded as medium probability of exposure, and kappa values for these comparisons were almost identical to those shown in table 2. The kappa values were recalculated after excluding those who were coded by the experts as having been exposed at a low concentration. Most of the kappa values remained about the same; the median kappa value was 0.54.

When the population controls and cancer cases were analyzed separately, they yielded very similar results. The median kappa was very similar for the two groups.

---

**TABLE 1.** Correspondence between terms used in self-completed questionnaire (SCQ) and terms used by experts

<table>
<thead>
<tr>
<th>SCQ term</th>
<th>Expert term(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fur or leather</td>
<td>Fur dust, leather dust</td>
</tr>
<tr>
<td>Wood or wood products</td>
<td>Wood dust</td>
</tr>
<tr>
<td>Glues or adhesive substances</td>
<td>Animal or vegetable glues; synthetic adhesives</td>
</tr>
<tr>
<td>Paint, wood stain, or varnish</td>
<td>Wood varnishes or stains; metal coatings; other paints and varnishes</td>
</tr>
<tr>
<td>Pesticides or fertilizers</td>
<td>Fertilizers; pesticides; DDT</td>
</tr>
<tr>
<td>Asbestos or insulation materials</td>
<td>Chrysotile asbestos; amphibole asbestos; mineral wool fibers; inorganic insulation dust; fiberglass</td>
</tr>
<tr>
<td>Pharmaceuticals</td>
<td>Pharmaceuticals</td>
</tr>
<tr>
<td>Lubricating oils and greases</td>
<td>Lubricating oils and greases</td>
</tr>
<tr>
<td>Gasoline or fuel oils</td>
<td>Leaded gasoline; diesel oil; aviation gasoline; jet fuel; kerosene; heating oil; crude oil</td>
</tr>
<tr>
<td>Solvents or degreasing liquids</td>
<td>Solvents</td>
</tr>
<tr>
<td>Plastic or rubber</td>
<td>Rubber dust; styrene-butadiene rubber; natural rubber; plastic dust; melanine-formaldehyde; urea-formaldehyde; phenol-formaldehyde; polyvinyl chloride; polyurethanes; polyesters; polyacrylates; polyamines; polyvinyl acetate; polyethylene; polypropylene; polystyrene</td>
</tr>
</tbody>
</table>

---

**TABLE 2.** Comparison of self-reports of occupational exposures with expert assessments, for each of 11 substances, among male subjects of a cancer study in Montreal, Quebec, Canada, 1979–1989

<table>
<thead>
<tr>
<th>Substance</th>
<th>2 x 2 table of concordance*</th>
<th>Indicies of concordance†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fur or leather</td>
<td>115</td>
<td>15</td>
</tr>
<tr>
<td>Wood or wood products</td>
<td>509</td>
<td>137</td>
</tr>
<tr>
<td>Glues or adhesive substances</td>
<td>350</td>
<td>114</td>
</tr>
<tr>
<td>Paint, wood stain, or varnish</td>
<td>340</td>
<td>203</td>
</tr>
<tr>
<td>Pesticides or fertilizers</td>
<td>109</td>
<td>79</td>
</tr>
<tr>
<td>Asbestos or insulation materials</td>
<td>294</td>
<td>262</td>
</tr>
<tr>
<td>Pharmaceuticals</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>Lubricating oils and greases</td>
<td>455</td>
<td>324</td>
</tr>
<tr>
<td>Gasoline or fuel oils</td>
<td>359</td>
<td>231</td>
</tr>
<tr>
<td>Solvents or degreasing liquids</td>
<td>612</td>
<td>434</td>
</tr>
<tr>
<td>Plastic or rubber</td>
<td>186</td>
<td>290</td>
</tr>
<tr>
<td>Median</td>
<td>0.51</td>
<td>0.61</td>
</tr>
</tbody>
</table>

* Number of subjects classified as exposed (E) and unexposed (U) by the subjects themselves and the experts (excluding subjects who were classed by the experts as exposed but with low confidence).
† The sensitivity and specificity are predicated on the assumption that the expert assessment is a "gold standard" for the self-report.
(0.57 for the population controls and 0.49 for the cancer cases). For nine of the 11 substances, the difference in kappa between the two groups was less than 0.10. Median specificity was almost identical for the two groups (0.89 and 0.90), and median sensitivity was slightly higher for the population controls (0.67) than for the cancer cases (0.61).

**DISCUSSION**

There are several methodological limitations to this study, which can be characterized as follows: data collection asymmetries; selection bias; and the issue of whether expert assessment really is an appropriate gold standard for evaluating self-assessment.

The data collection procedures differed in some crucial ways between the self-respondents and the experts. First, we used lay terms in the self-response questions, and there were some problems in meshing these terms with the technical terms used by the experts. Some of the misclassification may have been due to different interpretations of the lay terms by the subjects and the experts. However, self-completed questionnaires are typically limited to substances with names that are likely to be recognized by the average worker. The substances we listed were chosen because we believed they would be among the common occupational exposures and among the most evident to the worker. The validity of self-reports for these 11 substances are likely to represent a best-case scenario. Other substances, about which workers are less likely to know (e.g., phenol-formaldehyde and styrene) would result in much lower validity. However, in etiologic research, it is precisely at such a detailed level that we must aim, since a general class of substances (such as solvents or pesticides) may well dilute the presence of a small number of carcinogens by a larger number of noncarcinogens.

Another difference in data collection was that the self-administered questionnaire contained a checklist of substances to which the response was limited to yes, no, or don’t know, while the experts, on a four-point scale, indicated their confidence that the exposure had occurred. For the comparison, therefore, we had to dichotomize the expert categories; this was done by combining high and medium probability of exposure and dropping the category of low probability of exposure. This makes an assumption about the responses given by subjects who felt that they had low or medium probability of exposure to the substance. However, the kappa values did not change markedly when we excluded those subjects whom the experts had coded as medium probability of exposure.

Similarly, the kappa values did not change markedly when we excluded those whom the experts had coded as low concentration of exposure. This suggests that the workers do not omit exposures on the grounds that the concentration of the substance was low.

Another relevant issue in data collection was that our self-response exposure checklist was not job specific. That is, the respondent was not asked to relate the exposures to a specific job in his career; rather, he was asked whether he had ever been exposed. Our experts coded exposures on a job-specific basis. For the purpose of this paper, we merged the expert-derived exposure from across the man’s jobs and compared ever exposed (self-assessed) with ever exposed (expert assessed). To elicit job-specific data in the self-administered questionnaire places a greater burden on the respondent, and as a result, it is seldom done. However, ever/never information is superficial and imprecise. If we had asked questions about the checklist on a job-specific basis and used the job rather than the man as the unit of observation, the sensitivity and specificity would almost certainly have been lower. A man could recall true exposure to leather, for instance, in one of his jobs, and that would be counted as a concordant response. If he had been asked about each of his two jobs, he might have correctly reported the exposure for one but incorrectly failed to report the exposure in the other job. In that case, he would score one concordant and one discordant report. Thus, this analysis overestimates the agreement that would be observed if subjects were asked to report exposures for each job.

The second class of methodological problems concerns selection bias. Because we did not make any special effort to have the subjects fill out their self-assessed exposure checklist, less than half of the subjects in our large case-control study completed the self-assessed exposure questions (although a work history was obtained in the interview). The participants were significantly different from nonparticipants in several ways (higher incomes, French-speaking, less likely to be a proxy respondent). However, some of these characteristics of the participants were likely to make the responses to the questionnaire more accurate and should have decreased the resultant misclassification.

A related selection bias problem concerns our criteria for including a subject in this study; namely, subjects were included only if they checked at least one box on the page that presented the checklist of substances. A person who considered himself unexposed to all substances was supposed to tick the “no” box beside each substance. It is likely, however, that some of the subjects whom we considered as nonrespondents because they did not check any boxes on the checklist were, in fact, subjects who considered them-

Am J Epidemiol Vol. 144, No. 5, 1996
selves unexposed to all the substances. We have some evidence to support this because the respondents had substantially higher prevalences of exposure according to our experts than did the nonrespondents. This selection bias probably led to artificially high exposure prevalences and may have slightly inflated sensitivity and deflated specificity, but the latter effect is unlikely to have been large.

The third methodological issue is whether the expert assessment should really be considered as an appropriate gold standard for assessing self-reports. Our expert-assessment method is a complex and expensive way of assessing exposure, but it is acknowledged to be the best alternative in a population-based retrospective study (6). There are few solid data on the validity of the expert assessment of past exposure. From our own data, there is evidence that the procedure is repeatable, with a kappa value of about 0.80 (Siemiatycki et al., Institut Armand-Frappier, manuscript submitted for publication).

While it is certainly true that the self-respondents have the advantage of having personally experienced the working environments, the experts have some significant advantages when it comes to assessing exposure. It is possible that some workers are aware of the main raw materials, the processes, and the end-products in their workplaces. However, many workers do not know even this (7), or they may know the substances with which they work by their trade names and have no idea of the components of each product. In addition, the fact of exposure is a more complex issue that involves some understanding of chemical intermediates, possible contaminants, and industrial hygiene measures. Furthermore, the operational definition of what constitutes exposure in an epidemiologic study is one that is not self-evident and requires some technical understanding. For instance, some workers may consider a brief, episodic, low-level exposure to wood dust enough to warrant being mentioned, while others may have a much higher threshold before they consider themselves exposed. Experts can adopt criteria to promote consistency that is impossible with self-reports. This requires a comparative perspective on different occupations and different workplaces, which the average worker does not have.

While it is conceivable that a self-report may be more valid than an expert opinion when the expert is unaware of the worker’s assessment, this is less likely when the expert opinion incorporates the information from the self-report. That is, the experts were aware of the self-reports and used their expert judgment to weigh the validity of those self-reports. We therefore believe that, in most instances, the expert rating was closer to the “truth” than the self-report. Even if the expert rating was itself an “alloyed gold standard” (8), it is nonetheless instructive to estimate how much less perfect is the self-report.

A related issue is whether we used a self-assessment procedure that was much less accurate than that which is usually used. The questionnaire checklist included only 13 substances. While it would be possible to lengthen the list somewhat, it is not at all practicable to assess hundreds of substances in a self-report format because it would not be well tolerated by study subjects. Self-assessed exposure can therefore only be used when interest is focused on a small number of substances. In fact, many studies have used even shorter checklists than ours (9, 10). In contrast, an expert-based approach can evaluate hundreds of exposures since the limiting factor is the expense of the experts. Second, we used a closed question rather than an open-ended one because we wanted to stimulate thought about possible occupational exposures. It has been shown that questions listing exposure agents elicit more responses than does free recall (11). One study found that free recall resulted in the reporting of only 2.6 percent of exposures judged likely by an industrial hygienist (12). In addition, the collapsing of the respondents’ entire work history, the use of broad groupings of substances, and the lack of guidance as to how to consider concentration or frequency of exposure are common to most other self-reports (6, 13, 14).

Notwithstanding the above limitations, we believe this data set provides useful data concerning the validity of self-reports. Compared with the expert assessment, the median kappa was 0.51, and the median sensitivity and specificity were 0.61 and 0.90, respectively. The ranges of sensitivities and specificities in our study were similar to those found in a comparison of self-and expert-assessed exposure to eight substances in the printing and plastics industries (13). Specificities similar to ours were also reported in a comparison of self-assessed exposures and actual measurements in lumber mills (11). Sensitivity of the responses to individual metals (such as cadmium and copper) tended to be very low, while those to composite metals (such as tungsten carbide and stellite) were high.

To assess the practical impact of the misclassification that arises from using self-reports as opposed to expert assessments, we carried out an exercise to estimate the sample size that would be required to detect a hypothetical risk under the alloyed gold standard conditions of expert assessment and then under the misclassified conditions of self-reporting.

To estimate the required sample size under alloyed gold standard conditions, we had to postulate relative risk, prevalence of exposure, and alpha and beta. We
assumed alpha to be 0.05 (one-sided) and beta to be 0.20. A relative risk of 1.5 was chosen because it is unlikely that broad categories of substances such as the ones we examined would have a stronger relation with any disease. Although a specific carcinogen within the group may have a strong effect, its effect would be diluted in the broad category. Note that this relative risk of 1.5 pertains to the risk estimable by the experts; the true relative risk would be higher. We chose two prevalence levels and carried out the exercise with both. We chose 5 percent as representing a fairly common exposure and 20 percent as representing a very common exposure (in our data set of 294 substances, the median prevalence of exposure was 3 percent). Under these conditions, we would need a sample size of 1,323 cases (and the same number of controls) when the prevalence was 5 percent and 419 cases when the prevalence was 20 percent.

We then calculated the effect of misclassification using the median sensitivity and specificity from table 2 (0.61 and 0.90, respectively). The effect of this misclassification is to change the relative risk and the prevalence estimates according to formulae derived by Flegal et al. (15). Using these formulae, we found the modified relative risk to be 1.1 when the original prevalence was 5 percent and 1.2 when the original prevalence was 20 percent. Using these numbers in the standard sample size formula leads to a required sample size of 10,053 cases (and the same number of controls) when the original prevalence was 5 percent and 1,536 cases when the original prevalence was 20 percent. These increases in the sample sizes required to detect a risk would increase the total cost of a study considerably, likely compensating for the savings achieved by using the cheaper self-reporting procedure or possibly rendering it altogether unfeasible.

In fact, these computations may be somewhat misleading. In the situation of adjusting an observed odds ratio for misclassification, Wacholder et al. (8) showed that when the estimates of sensitivity and specificity are derived from comparisons with an alloyed gold standard, the adjustment procedure is not necessarily valid. While our procedure of sample size estimation was not identical to the one they analyzed, it appears to be analogous, and we therefore expect that their conclusion would hold. Thus, the sample size requirement under misclassification could be greater or less than that computed above, depending on the unknowable correlations between the two types of exposure assessment and the truth. Imperfect though they may be, these estimates provide some guidelines as to the possible impact of such misclassification.

As in other studies (16, 17), we found no important differences in the accuracy of self-reported exposures between population controls and cancer cases. This suggests that any recall bias that may be present in self-reports of exposure is non-differential with respect to case or control status, and this is reassuring with regard to the more general question of differential reliability of occupational history.

In conclusion, this study demonstrates that the use of typical self-reports of exposure in population-based studies of occupational risk factors entails considerable misclassification of exposure compared with expert assessment. There would be even greater misclassification if we had asked about relatively esoteric substances such as naphthalene or benzo(a)pyrene. As a general rule, we do not believe that self-reports are sufficiently accurate to warrant their sole use in population-based studies, although in special situations their use might be justifiable. We do, however, believe that self-reports can be a useful component of a broader exposure assessment strategy.

ACKNOWLEDGMENTS

This study was funded by grants from the Institut de recherche en santé et en sécurité du travail du Québec, the Fonds de recherche en santé du Québec, the National Health Research and Development Program, and the National Cancer Institute of Canada.

The authors are grateful to Dr. Louise Nadon and Marie Désy for advice on the expert coding and the statistical methods, respectively, and to Dr. Patricia Stewart for comments on an earlier draft of this paper. Important contributors to the study included Denis Bégin and Ramzan Lakhani.

REFERENCES


