Ann. occup. Hyg., Vol. 47, No. 6, pp. 461–475, 2003
© 2003 British Occupational Hygiene Society
Published by Oxford University Press
DOI: 10.1093/annhyg/meg066

Expert Judgment and Occupational Hygiene: Application to Aerosol Speciation in the Nickel Primary Production Industry

GURUMURTHY RAMACHANDRAN[1]*, SUDIPTO BANERJEE[2] and JAMES H. VINCENT[3]

[1]Division of Environmental and Occupational Health, School of Public Health, University of Minnesota, Minneapolis, MN 55455; [2]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455; [3]Department of Environmental Health Sciences, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA

Received 4 December 2002; in final form 23 April 2003

In many situations characterized by sparse data, occupational hygienists have used subjective judgments that are claimed to be derived from their experience and knowledge. While this practice is widespread, there has been no systematic study of 'expert judgment' or the 'art' of occupational hygiene. Indeed, there is a need to address the question of whether there is such a thing as 'expert opinion' in occupational hygiene that is broadly shared by practicing professionals. This research, employing 11 experts who estimate an exposure parameter (the percentages of four nickel species) in 12 workplaces in a nickel primary production industry, provides a large dataset from which useful inferences can be drawn about the quality of expert judgments and the variability among the experts. A well-designed questionnaire that provided succinct information about the processes and baseline data served to calibrate the experts.

The Bayesian framework has been used in this work to develop posterior means and standard deviations of the percentages of the four nickel species in the 12 workplaces of interest in the company. These estimates of the nickel speciation are at least as precise as—and most of the time more precise than—those provided by the sparse measurement data. There was a very high degree of agreement among the experts. A majority of the experts agreed among themselves 92% of the time, while almost two-thirds agreed 73% of the time. This, coupled with the fact that the experts came from varied backgrounds, seems to suggest that there is indeed some broad body of specialized knowledge that the experts are drawing on to reach similar judgments.

It also seems that one type of expert is not necessarily any better than any other kind, and expertise does not necessarily require intimate familiarity with the workplace. In this example, the expert judgment exercise has indeed enhanced the quality of our knowledge of the exposure 'fingerprints' for the nickel industry workplaces studied and the combination of expert judgment and sparse data is better than the sparse data alone. For occupational hygiene exposure assessment, our experience suggests that such expert judgment methods can provide a cost-effective means to improve and refine information about workplace hazards. However, more study is warranted for situations where the domain of the quantity of interest has a much wider range of values, e.g. actual exposure values.

Keywords: Bayesian framework; expert judgment; exposure assessment; nickel speciation; sparse data

INTRODUCTION

Occupational hygiene is defined as the practice of the science and art involved in the anticipation, recognition, evaluation and control of hazards in the workplace. Exposure assessment is an important part, embracing the first three of these four elements. It

*To whom correspondence should be addressed. Tel: +1-612-626-5428; fax: +1-612-626-0650; e-mail: ramac002@umn.edu

provides the knowledge that is essential to determine the presence of a hazard and to quantify it. Determining a relationship between some measure of exposure with some observed plausible adverse health effect can enable the establishment of scientifically based standards. In day-to-day occupational hygiene practice, the determination of exposure is a vital part of the assessment of compliance with either regulatory or voluntary standards.

Modern occupational hygienists recognize that exposures to, say, airborne chemical contaminants are very variable. In addition, exposure assessment may be required for several chemical species simultaneously. For example, in nickel exposure assessment, the American Conference of Governmental Industrial Hygienists (ACGIH, 2001) specifies TLVs for a number of nickel species, including (i) elemental/metallic, (ii) soluble, (iii) insoluble and (iv) subsulfide, which may be present together in a given workplace. Thus, an adequate exposure assessment that captures this variability for multiple species can be very time-consuming and costly. The cost mounts if the analysis of the individual samples requires sophisticated instrumentation and/or analytical methods to identify specific health-related components or species (Vincent *et al.*, 1995, 2001).

For these reasons, it is frequently the case that insufficient exposure data are available to assess risk by conventional occupational hygiene scientific methods. Therefore, professional occupational hygienists routinely interpret such data as are available by informally using their own ‘expert judgment’, claiming that this, in effect, increasingly invokes that part of occupational hygiene often referred to as ‘art’. While, in the past, this has been regarded as something that is intangible and arising from their professional experiences, occupational hygiene researchers are beginning to avail of a formal body of methodology known as ‘expert judgment science’ that has worked well in other fields (e.g. Kromhout *et al.*, 1987; Hawkins and Evans, 1989; Post *et al.*, 1991; Cherrie and Schneider, 1999; Ramachandran, 2001). In these previous studies, occupational hygiene experts were used to obtain estimates of inputs to deterministic exposure models (e.g. inputs to the general ventilation model) that were independent of available exposure measurements. The output of the exposure model was the exposure estimate (along with the uncertainty in the estimate). While the experts appeared to agree with each other, these studies employed a very limited number of experts, and the variability in expert opinions was not studied. Additionally, the experts used in those studies (e.g. Ramachandran, 2001) came from very similar professional backgrounds (academic scientists with considerable expertise in aerosol science in occupational settings, although one expert had considerably more experience with the nickel industry) and so it was not clear whether the convergence of expert opinions was merely a reflection of this fact. In short, these studies did not address the question of whether there was such a thing as ‘expert opinion’ in occupational hygiene that was broadly shared by practicing professionals.

This paper is a contribution to an emerging body of work dealing with expert judgment in occupational hygiene. It describes the application of Bayesian ideas to the comparison of expert opinions, mathematically combining expert opinions and refining these combined expert opinions with actual workplace measurements. Eleven experts from varied backgrounds in occupational hygiene were provided with a description of the processes that occur in a primary nickel production worksite to convert nickel ore to metallic nickel as well as results from previous studies of nickel speciation that act as anchoring information and helps the experts to calibrate their responses. These experts were then asked to estimate the percentages of four different nickel species in 12 different worksites within the company. Each set of percentages of the four different nickel species is called a ‘fingerprint’. Thus, each of the 11 experts had to provide 48 expert opinions in the form of probability distributions—this is a sufficiently large dataset with which to analyze the variability in expert opinions. This study is built around an extension of the field study of Vincent *et al.* (2001) in the nickel primary production industry that had been aimed at developing ‘fingerprints’ for aerosol exposures for the purpose of facilitating future routine exposure assessment. Such ‘fingerprints’, if they were reasonably stable, would then serve to allow the results of routine exposure assessments (i.e. carried out simply with respect to total nickel, as is done at present) to be used to infer information about nickel species content at individual worksites. One of the issues was the extent to which meaningful conclusions could be drawn from the quite sparse measured data that had been obtained from what had nonetheless been a very expensive and time-consuming field study. The methods of expert judgment could possibly play a useful role in this context. In this paper, therefore, we explore that role, using the workplace nickel speciation as a concrete and meaningful example of how formalized expert judgment methods might be useful more widely in occupational hygiene exposure assessment.

BASIC APPROACH

Rationale

As already mentioned, the study derived from previous field research in which ‘fingerprints’ were sought for people’s exposures to various nickel-containing species in primary nickel industry producing worksites. That is, for each worksite, measurements were made of the distribution of nickel species within the inhalable particle size fraction, yielding the percentages of water-soluble, sulfidic, oxidic and metallic species groups (where all these percentages add up to 100%).

We set out to explore how the limited amount of data in that earlier research may be refined by knowledge about the distributions of nickel species for the various worksites reflected in the body of informed opinion drawn from a carefully chosen panel of well-qualified experts. By combining the actual worksite data on nickel speciation with corresponding estimated data obtained from the panel, we set out to obtain refined exposure 'fingerprints' with defined and acceptable variability. This was achieved using the Bayesian methodology that has already been applied elsewhere in the formalized analysis of expert opinions (e.g. Cooke, 1991).

Mathematical basis

In the Bayesian approach, actual physical measurement of a quantity of interest serves to refine previous knowledge of that quantity by adjusting its probability distribution. It is thus based on inductive reasoning. This process of reasoning is very close to the application of the 'art' of occupational hygiene, where it is common for practitioners to make initial educated guesses about workplace exposures (e.g. even if only at the crude level of whether exposures are high or low, or containing mainly one chemical species) and then to refine (i.e. confirm or deny) it by making actual measurements of exposures.

If the physical quantity of interest is represented by f, and the measured data are represented by m, then the updated probability distribution of f is given by

$$P_{\text{posterior}}(f/m) = \frac{P_{\text{prior}}(f)P_{\text{L}}(m/f)}{P(m)} \qquad (1)$$

Here $P_{\text{prior}}(f)$ is the probability distribution of f prior to making any measurements (the 'prior'), while $P_{\text{L}}(m/f)$ is the likelihood that, given the true value f, the measurement m is observed. In addition, $P(m)$ is the probability that the measurement m is observed, and $P_{\text{posterior}}(f/m)$ is the updated probability that the physical quantity of interest is f, given that measurements m are observed (the 'posterior') (Makridakis *et al.*, 1984; Little and Rubin, 1987). For application in the example chosen for this research, f is the proportion of a given nickel species found at a given worksite, and m is the measured value of f. In this equation, the important terms on the right-hand side are those in the numerator, and the denominator $P(m)$ serves simply to normalize the expression so that the integral of $P_{\text{posterior}}(f/m)$ over all f (from 0 to 100%) is unity. Ideally, the updated probability will provide a narrower probability distribution of (and hence greater confidence in) the quantity of interest than either the subjective initial probability provided by the experts or the objective (but incomplete and uncertain) measurements taken individually.

Measured data

In an earlier exposure assessment study, nickel speciation was conducted for a proportion of the personal inhalable aerosol samples collected at the worksites visited, not only in the primary nickel production industry but also in the nickel alloy production and nickel electroplating industries (Vincent *et al.*, 1995). Quantitation of samples with regard to the four species groups was performed using the sequential extraction procedure of Zatka *et al.* (1992). The results showed that the appearance of the fractional nickel content of water-soluble, sulfidic, oxidic and metallic species groups varied greatly from one plant to another throughout the industry.

For the present research, the results of the earliest study (Vincent *et al.*, 1995), presented in the form of plant-wide averages for the mining/milling, smelting and refining plants (for just the primary nickel production industry), were used to provide 'anchoring' or 'calibration' information for the experts (see Table 1). The results of the second study (Vincent *et al.*, 2001), for a range of individual worksites in milling, smelting and refining plants, were the ones of primary interest here in relation to the 'fingerprinting' exercise, and so these data were not revealed to the experts. These data are shown in Table 2.

In the 2001 study, sampling was carried out using an Andersen cascade impactor, modified by the addition of a top stage comprising a coarse foam filter plug in order to extend the range of the instrument over a large proportion of the inhalable range (i.e. up to particle aerodynamic diameters of close to 100 μm). In addition, inhalable aerosol samples were collected using the Institute of Occupational Medicine (IOM) personal inhalable aerosol sampler. The IOM samplers were mounted on life-size mannequins to simulate their being worn by actual workers. From the outset we acknowledged the possibility that the distribution of nickel species may not be the same across the whole range of inhalable particle sizes, and this was the reason for the choice of the Andersen sampler. As it transpired, however, no significant differences in

Table 1. Summary of nickel speciation data from the first study that was provided to the experts (Vincent *et al.*, 1995) for calibration purposes (all data are percentages)

	Mining (inc. milling)	Smelting	Refining
Soluble	20 (15–27)	9 (6–12)	5 (3–7)
Sulfidic	45 (40–45)	44 (30–67)	7 (3–15)
Metallic	4 (0–8)	5 (3–8)	12 (5–15)
Oxidic	32 (27–34)	32 (7–50)	72 (60–81)

Table 2. Summary of results for nickel species groups at the worksites indicated, from pooled modified Andersen and IOM sampler results (from Vincent *et al.*, 2001); these were used as measurement data for Bayesian updating

Company and process	Worksite	No. of samples	Average percentage (SD)			
			% Soluble	% Sulfidic	% Metallic	% Oxidic
Mill	Tipple	4	5.8 (1.8)	61.8 (4.2)	8.1 (3.6)	24.3 (2.3)
	Grinding	4	7.9 (1.5)	50.3 (19.4)	7.7 (5.6)	34.1 (18.7)
Smelter	Flash furnace	3	3.8 (0.5)	69.2 (12.0)	9.1 (9.6)	17.9 (4.7)
	Converter aisle	4	8.0 (4.7)	34.8 (19.8)	13.5 (5.6)	43.6 (19.9)
	Matte crushing	4	4.3 (2.8)	73.3 (12.0)	13.6 (8.6)	8.9 (8.4)
	Matte processing	6	3.6 (2.2)	67.8 (13.8)	12.2 (6.5)	16.5 (10.1)
	FBR	6	4.3 (1.3)	24.4 (16.3)	4.0 (5.2)	67.4 (21.3)
	Cottrell	2	8.5 (3.2)	22.0 (8.7)	5.5 (3.3)	64.0 (2.2)
Refinery	Feed receiving	4	2.0 (3.1)	10.1 (7.7)	10.0 (10.7)	77.9 (19.3)
	TBRC	4	6.7 (0.8)	9.2 (5.9)	6.7 (1.9)	77.3 (4.5)
	Ball mill	2	2.6 (3.7)	14.0 (11.6)	32.1 (29.7)	51.2 (21.8)
	Packaging	6	1.5 (1.3)	4.1 (2.5)	31.9 (11.7)	62.7 (10.6)

FBR, fluidized bed roaster; Cottrell, electrostatic precipitator area; TBRC, top-blown rotary converters.

the measured results were observed as a function of particle size. This finding allowed expression of the desired 'fingerprints' simply in terms of the distributions of the nickel species alone. In turn, this enabled pooling of the Andersen and IOM sampler results in order to provide a larger data set. Despite the pooling of data, the number of data points per worksite was still quite limited.

Table 2 indicates the distribution of the four nickel species groups in various worksites. These results show differences from worksite to worksite. For some worksites, notably the mills, differences from worksite to worksite were found to be not statistically significant whereas statistically significant worksite-to-worksite differences were found for the smelter and the refinery (Vincent *et al.*, 2001). In general, the observed speciation differences were qualitatively consistent with what is known about the metallurgical processes taking place throughout the worksites in question.

MATERIALS AND METHODS

Design of the questionnaire

A questionnaire was developed in order to elicit opinions from selected experts about the distributions of the four nickel species groups for each of the worksites reported in the 2002 study. It set out to meet the following basic specifications:

- to be pitched at a level commensurate with the expertise of the selected experts;
- to contain full but concise information about the processes taking place at all the worksites studied in the worksites of interest (developed during the visits to the worksites and in consultation with the professional staff at those worksites);
- to contain sufficient relevant data in order to provide the experts with baseline information about nickel species in the worksites of interest (as was available from the 1995 speciation study).

In designing the questionnaire document, it was acknowledged that the chosen experts were very busy people. So the information that was being sought would need to be very concise, clearly identified and articulated without ambiguity. With this in mind, the document that was sent to each expert took the form of a letter that contained a brief summary of the rationale for the study, short paragraphs summarizing the nature of the metallurgical processes taking place at each of the worksites studied in the field campaign, summary data from the 1995 field study (as shown in Table 1), and a clear statement of what the expert is being asked to provide. The expert was asked to give his/her opinion about the proportion of each nickel species group at each worksite, expressed as the 95% confidence interval of a probability distribution. The expert was informed that the limits of the confidence interval would be interpreted in subsequent analysis as the 2.5th and 97.5th percentiles of a uniform probability distribution. Thus, if for a given worksite and a given nickel species an expert gave the range as 20–40%, then it implies that the expert is 95% confident that the true percentage falls in this range (with all values being equally likely), and 5% confident that the true value lies in the combined ranges 0–20% and 40–100%.

Choice of experts

Experts were chosen from one of the following three categories:

1. Primarily aerosol scientists with occupational hygiene expertise in non-nickel industries.
2. Primarily occupational hygienists with experience in aerosol exposure assessment.
3. Occupational hygiene scientists with specific knowledge of the nickel primary production industry.

Twenty-five such individuals were identified and contacted, and initially agreed to participate. They were all sent the questionnaire and asked to respond within a specified time frame. Eleven completed responses were received within the time frame allotted, and these experts are identified in this paper as experts A–K. They come from five different countries, and the list includes highly respected individuals at the highest levels in the fields indicated. Experts B–D fall into category 1; experts A, F–I fall into category 2; and experts E, J and H fall into category 3.

Calculating a posterior using an expert prior and the measured data

Prior information, reflected in the form of probability distributions, plays an integral part in Bayesian statistical inference. Bayesian inference proceeds by updating this prior information by incorporating fresh information from collected data through Bayes's theorem. The experts use the process description and the limited historical data that were provided to them to form the prior distributions. But the fresh information referred to should be based on new data, and not the same data that the experts used to form their prior distributions. The result is a posterior distribution for the parameters, and the Bayesian philosophy uses this posterior distribution to carry out all inferences.

In our research, the experts expressed their prior beliefs in the form of uniform distributions. On the other hand, the measured percentages of the four nickel species are assumed to be normally distributed with mean μ and standard deviation σ. Usually, statistical inference is desired for one or both of these parameters. However, since our objective is the study of variation in prior beliefs and not actual complete Bayesian inference (there are too few data anyway), we assume that the parameter of interest is the population mean μ.

Thus, each expert has an associated 'belief distribution', say $P(\mu|a,b)$, for the mean percentage (μ) of a particular nickel species in the population. Here a and b are parameters, usually specified by the expert, that control the nature of his/her prior belief. In the present context, each expert summarizes his/her belief about the mean percentage (μ) in terms of a uniform distribution—they specify an interval within which the percentage mean is supposed to lie. The shorter the length of this interval, the more precise the expert is about his prior belief; the longer it is, the less precise he/she is. Thus, a and b will denote the left and right end-points of the expert's belief. This prior belief is to be updated with the measured data using Bayes's theorem, and expressed in the form of a posterior distribution $P(\mu|a,b,$ sample data) (see Appendix).

Combining expert opinions

As the first step in the analysis of the data received from the experts, each percentage range reported for a given species in a given worksite was expressed graphically as a uniform distribution, with 95% of its area contained with the range indicated. The remaining 5% was distributed uniformly over the whole of the rest of the range from 0 to 100% falling outside the 95% interval. In order to illustrate the method for reducing the data from the experts' responses, Fig. 1 shows an example for just two experts. Here it is important to note that the total areas under the graphs for the two experts must be the same. These two distributions may now be added, and the result—also as shown in the figure—is a new distribution based on the combined opinion of both experts. The corresponding result for the combined opinions of all 11 experts is obtained in the same manner.

RESULTS AND DISCUSSION

Comparing experts

Figure 2 shows the prior beliefs of all 11 experts expressed as the upper and lower 95% percentiles of a uniform distribution for one worksite (the smelter matte crushing area). Due to limited space, we will not show the expert priors for all 12 worksites. For the purposes of this paper, we will consider two experts to be in agreement if their 95% confidence intervals overlap. We notice that on no occasion is there total unanimity among the experts (i.e. for each of the four nickel species, there are at least two experts whose 95% confidence intervals for the percentage nickel species do not overlap). Table 3 summarizes, for all worksites, the agreement between experts and is based only upon prior inference. For every site-species combination (i.e. for every cell), we assign '1' if the experts failed to be unanimous and we assign '0' if they were unanimous. Note that, barring one cell, all the other cells reveal significant variation among the expert beliefs. However, this is an unduly pessimistic assessment of agreement between the experts. It obscures the degree of agreement between experts and assigns a '1' even if 10 of the 11 experts were in agreement. We can define a quantity called the *fractional agreement* as the number of experts whose beliefs overlap with each other. In Fig. 2, this can be determined by drawing a horizontal straight line such that it intersects the maximum number of expert opinion distributions. For example, in Fig. 2 for soluble nickel, except for

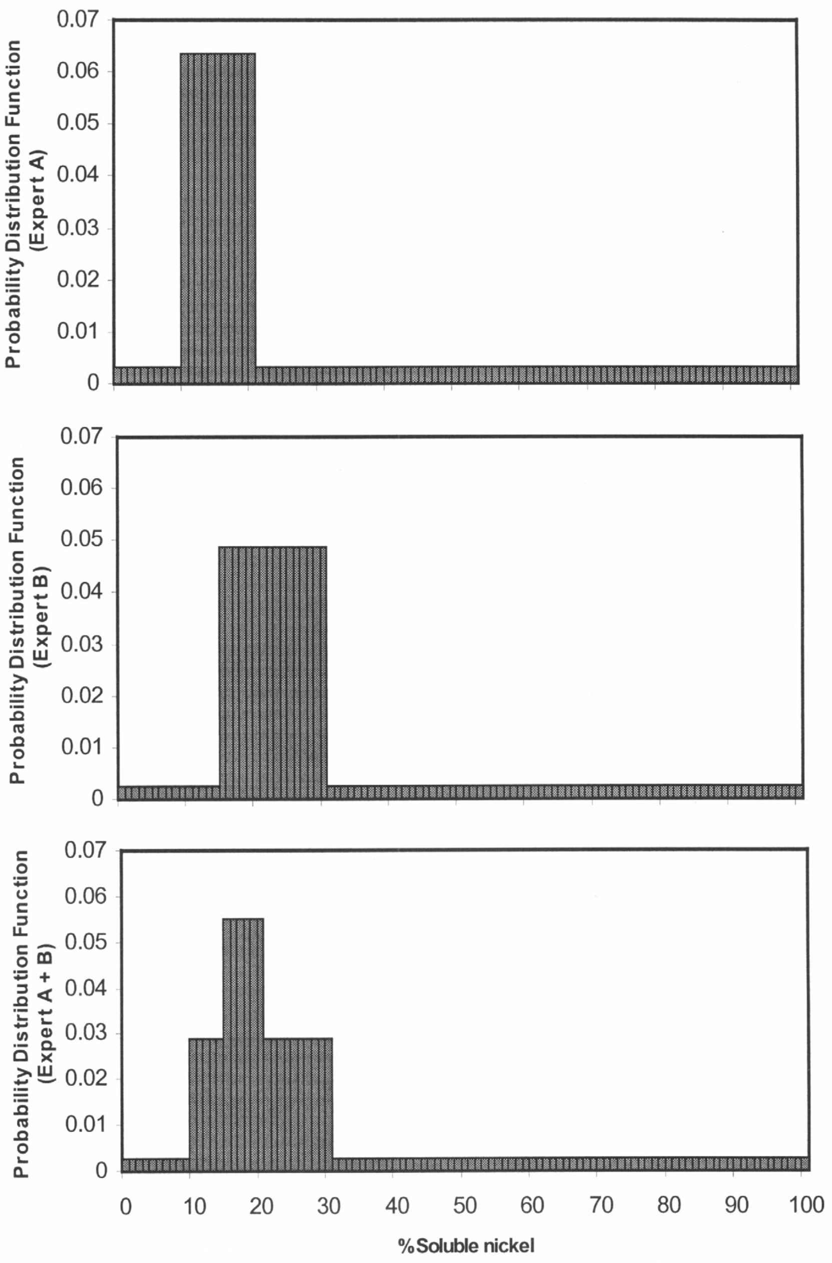

Figure 1. Example illustrating how the probability distributions of two experts are used to obtain a consensus (average) expert judgment.

experts I and K, all the others are in agreement so that the fractional agreement is 9/11 = 0.82. For metallic nickel, with the exception of experts A, B and D, all the others are in agreement, and the fractional agreement is 8/11 = 0.73. For oxidic and sulfidic nickel, the fractional agreements are also equal to 0.73. Thus for each of the four situations in Fig. 2, there is a high degree of agreement between the experts. To reflect this, Table 3 also shows the fractional agreement between the experts in parentheses. Out of the 48 situations shown in the table, a majority of the experts (6 or more out of 11) agree in 44 situations (i.e. 92% of the time). Seven or more experts (i.e. almost two-thirds) agree in 35 situations (i.e. 73% of the time).

The above analysis compared only the prior distributions provided by the experts, without reference to the measurements. However, in a Bayesian framework, the final decision is based on the posterior distribution. One can imagine a situation where two expert priors may be different in a statistically significant manner but the variability in the measurements may be much greater than the variability between the experts. In this case, the two expert posteriors will not

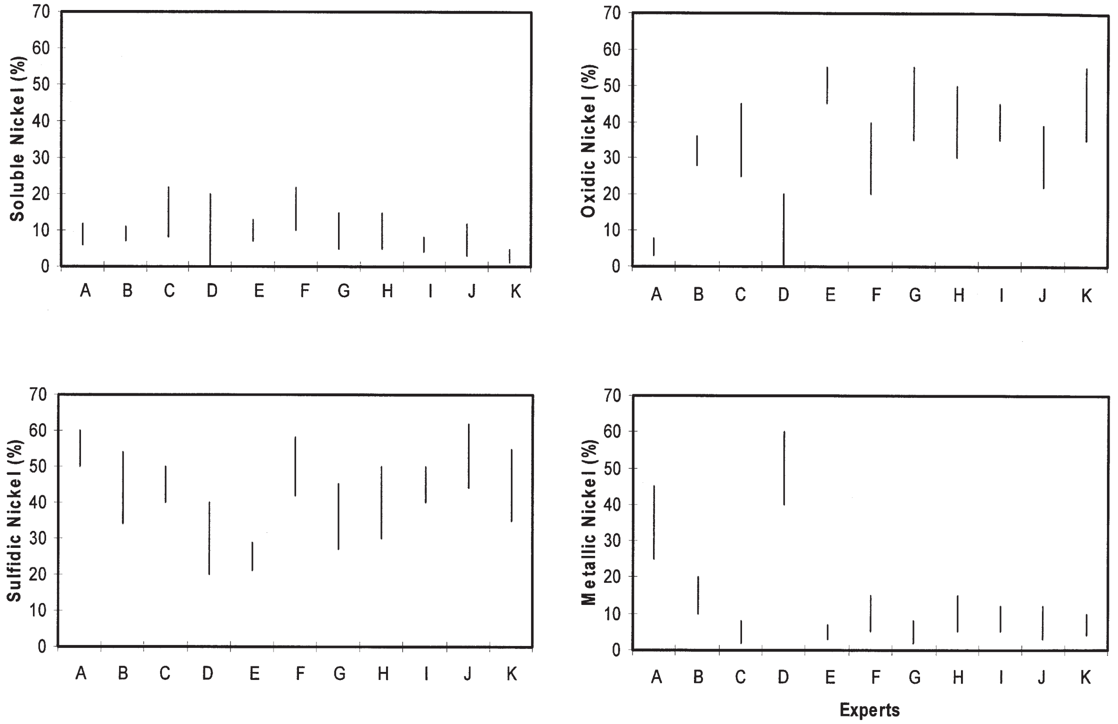

Figure 2. Priors of 11 experts for four nickel species in the smelter matte crushing area. Each prior is a uniform probability distribution between the limits shown in the graph.

Table 3. Agreement between expert priors for all worksites and all nickel species

Site	Soluble	Sulfidic	Oxidic	Metallic
Mill tipple	**1** (0.91)	**1** (0.91)	**1** (0.64)	**1** (0.82)
Mill grinding	**1** (0.82)	**1** (0.82)	**1** (0.64)	**1** (0.73)
Smelter flash furnace	**1** (0.82)	**1** (0.64)	**1** (0.55)	**1** (0.82)
Smelter converter aisle	**0** (1.00)	**1** (0.73)	**1** (0.73)	**1** (0.82)
Smelter matte crush	**1** (0.82)	**1** (0.73)	**1** (0.73)	**1** (0.73)
Smelter matte process	**1** (0.82)	**1** (0.64)	**1** (0.55)	**1** (0.55)
Smelter FBR	**1** (0.91)	**1** (0.45)	**1** (0.55)	**1** (0.55)
Smelter Cottrell	**1** (0.64)	**1** (0.55)	**1** (0.73)	**1** (0.64)
Refinery feed receiving	**1** (0.64)	**1** (0.45)	**1** (0.36)	**1** (0.64)
Refinery TBRC	**1** (0.82)	**1** (0.55)	**1** (0.64)	**1** (0.45)
Refinery ball mill	**1** (0.82)	**1** (0.73)	**1** (0.73)	**1** (0.55)
Refinery packaging	**1** (0.73)	**1** (0.73)	**1** (0.64)	**1** (0.64)

For every worksite and species combination (i.e. for every cell) we assign '1' if the experts had failed to be unanimous and we assign '0' if they were unanimous (the fractional agreement between the 11 experts is shown in parentheses).

be statistically different from each other. In a converse situation, the two priors may not be statistically different, but the measurements may have a high precision such that the posteriors will be statistically different. Therefore, in order to analyze variation in beliefs among the experts, it would be sensible to say that two experts differ significantly from one another in their opinions if the two posterior distributions arising from their respective priors are significantly different.

With this in mind, therefore, we adopted the following strategy to study posterior variation. We simulated data from the truncated normal distributions (the posteriors) associated with each of the 11 experts. We then created box-plots of the central 95% of these distributions, where (see Fig. 3a,b for examples) the line inside the box is the median, the ends of the box represent the interquartile range and the whiskers represent the 95% confidence intervals. If the box-plots overlapped, then we concluded that

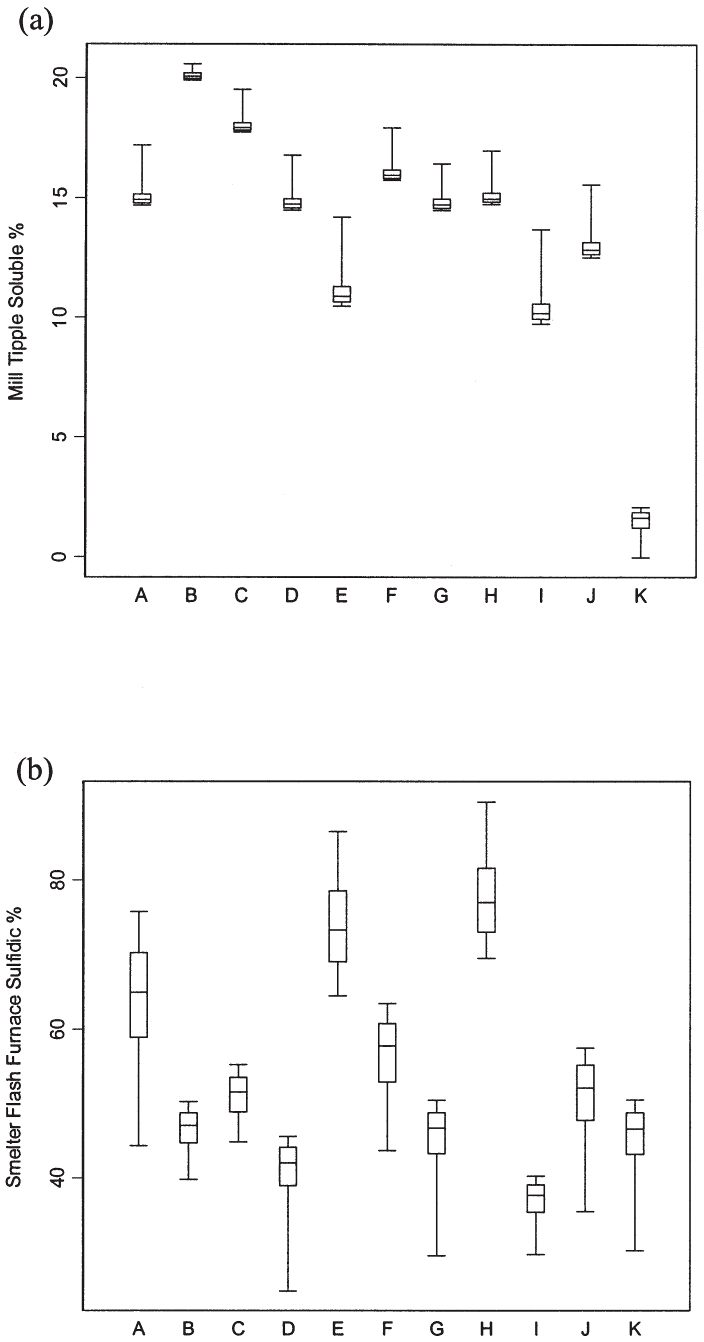

Figure 3. (a) Posterior distributions of soluble nickel for 11 experts in the mill tipple area. (b) Posterior distributions of sulfidic nickel for 11 experts in the smelter flash furnace.

even if the experts' prior beliefs were different, the differences were not significant enough to reflect in their posterior distributions. On the other hand, if the box-plots did not overlap, the posterior distributions were considered to be significantly different, leading to significantly different conclusions. In particular, here we would conclude that the experts truly do have different beliefs. Typically, more data means more updating information and this diminishes the impact of the priors on the inference. However, when such is not the case—i.e. when data are sparse—the priors tend to have a far greater impact upon the inference. Figure 3a,b show the posterior beliefs of all 11 experts for two worksites (the mill tipple and the smelter flash furnace areas) and all species expressed as box- and whisker-plots.

Table 4. Agreement between expert posteriors for all worksites and all nickel species

Site	Soluble	Sulfidic	Oxidic	Metallic
Mill tipple	**1** (0.45)	**1** (0.73)	**1** (0.64)	**1** (0.73)
Mill grinding	**1** (0.45)	**1** (0.82)	**1** (0.64)	**1** (0.73)
Smelter flash furnace	**1** (0.82)	**1** (0.73)	**1** (0.55)	**1** (0.82)
Smelter converter aisle	**0** (1.00)	**1** (0.82)	**1** (0.82)	**1** (0.91)
Smelter matte crush	**1** (0.82)	**1** (0.73)	**1** (0.73)	**1** (0.73)
Smelter matte process	**1** (0.91)	**1** (0.73)	**1** (0.55)	**1** (0.64)
Smelter FBR	**1** (0.73)	**1** (0.55)	**1** (0.55)	**1** (0.64)
Smelter Cottrell	**1** (0.91)	**1** (0.64)	**1** (0.36)	**1** (0.82)
Refinery feed receiving	**1** (0.73)	**1** (0.45)	**1** (0.45)	**1** (0.64)
Refinery TBRC	**1** (0.92)	**1** (0.64)	**1** (0.64)	**1** (0.64)
Refinery ball mill	**1** (0.82)	**1** (0.92)	**1** (0.82)	**1** (0.64)
Refinery packaging	**1** (0.82)	**1** (0.73)	**1** (0.73)	**1** (0.73)

For every worksite and species combination (i.e. for every cell) we assign ‘1’ if the experts had failed to be unanimous and we assign ‘0’ if they were unanimous (the fractional agreement between the 11 experts is shown in parentheses).

Table 5. Matrix of fractional agreement between experts

	A	B	C	D	E	F	G	H	I	J	K
A		0.48	0.46	0.67	0.48	0.38	0.46	0.52	0.38	0.46	0.25
B			0.83	0.60	0.52	0.65	0.73	0.67	0.65	0.79	0.48
C				0.65	0.77	0.90	0.92	0.73	0.77	0.88	0.71
D					0.73	0.69	0.67	0.44	0.58	0.65	0.54
E						0.63	0.77	0.56	0.63	0.71	0.50
F							0.79	0.58	0.65	0.79	0.56
G								0.71	0.88	0.94	0.69
H									0.69	0.65	0.52
I										0.83	0.58
J											0.79
K											

Table 4 is based upon posterior inference. Similar to Table 3, for every site species combination (i.e. for every cell) we assign ‘1’ if the experts had failed to be unanimous and we assign ‘0’ if they were unanimous. Again, barring one cell, the expert opinions always produced significantly different posteriors. In fact, there is no difference in the patterns of 1s and 0s in Tables 4 and 5. However, the fractional agreement between the experts is quite different for the posteriors than for the priors. A comparison of the two tables also shows that, in some instances, the agreement between the experts is better for the priors than for the posteriors. In other instances the reverse is true.

An examination of why this happens is revealing. In the case of soluble nickel in the mill tipple, the agreement between experts for the priors is 91%, whereas for the posteriors it is 45%. This is because for 10 of the 11 experts, the uniform priors range between 10 and 35% whereas the actual measurements indicate a mean of 5.8% with a rather narrow standard deviation of 1.8%. Since the measurements are quite different from the priors, the resulting posteriors are truncated normal distributions that are also narrower than the priors (see Fig. 3a). This leads to more expert posteriors differing significantly from each other. Interestingly, in this instance, only expert K was in agreement with the measurements.

In the case of sulfidic nickel in the smelter flash furnace area, the agreement between priors is 64%, whereas the agreement between the posteriors is 73%. The actual measurements had a mean of 69.2% and a standard deviation of 12%. So, for the priors, four of the experts differed significantly from the rest. One of them (expert D) had a prior range from 25 to 45%. The corresponding posterior had a slightly higher range extending beyond 45%. At the same time, the posteriors of the other experts were spread out further to lower values (see Fig. 3b). This resulted in the posterior of expert D being not significantly different from seven other experts, and thus eight experts were now in agreement. In this example, the sparse data exhibited greater variability than the differences between the priors, and thus the posteriors were not significantly different.

The effect of the priors on the data may be explained as follows. Typically, the data will try to move the prior mean towards the sample mean, thereby yielding a posterior mean that is a weighted average of the two. In exact normal inference (with normal priors and data) we know that these weights are the relative precisions (Carlin and Louis, 1998). Thus, the posterior mean shrinks towards the data mean if the precision of the data is higher (i.e. the variance is lower) than the precision of the prior and it shrinks towards the prior mean if the precision of the prior is more than that of the data. Even in our setting (with uniform priors) the above dynamics hold, as is evident from the means in Table 6. However, it cannot bring the prior mean beyond the support (or range) of the prior distribution. That is, the posterior mean (the 'final' Bayesian estimate of the mean) must lie within the interval $[a,b]$ specified by the experts. Here, the priors are uniform and the data is normal. This yields truncated normal posteriors so that the data, while trying to bring down the means, cannot do so below the lower endpoint of the prior. Note that if the experts had summarized their beliefs using normal priors, we would have obtained untruncated normal posteriors. This would have moved the posterior means towards the mean of the data without the restriction of the interval provided by the experts. However, that is not the case here.

Specifically, let us look at the priors for percentage of soluble nickel, corresponding to experts A and B, for mill-tipple. The experts quantify their priors using the 2.5 and 97.5% percentiles for a uniform distribution. The first one is (15,27) and the second one is (20,24). This is equivalent to specifying uniform distributions with end-points ($a = 14.68$, $b = 27.31$) and ($a = 19.89$, $b = 24.11$), respectively. The mean from the measured data is 5.8. Both the posterior means try to go towards 5.8. But the first cannot go below 14.68 and the second cannot go below 19.89 as demanded by the priors. Since 14.68 and 19.89 are different, the 'centers' of the resulting posterior distributions become separated. In fact, the posterior (2.5%, 97.5%) percentiles obtained for experts A and B are, respectively, (14.69, 15.87) and (19.90, 20.58). Note that both the posterior intervals are far tighter than their prior counterparts. This is precisely the shrinking effect of the data mean, which is tightly centered about 5.8 with a standard error of 1.8.

Defining consensus

Table 5 shows a matrix where each element is the fractional agreement between any two experts. Thus, a value of 0.65 for agreement between experts B and F means that out the total of 12 work sites × 4 species = 48 instances, the two experts agreed on 31 instances (i.e. 69% of the time). We can use a matrix such as this to identify individual experts who tend to disagree with the rest most of the time (their fractional agreement is <0.5). Using this somewhat arbitrary criterion, we see that expert A typically disagrees with the other experts. To a lesser extent this also true of expert K. In a similar manner, we can determine which experts are most in agreement with other experts. Again, if we arbitrarily set a minimum limit of 0.65, we find that experts C, G and J agree with nine other experts at least 65% of the time. It is emphasized that this analysis does not imply that expert A is necessarily wrong or that C, G and J are necessarily 'good' experts. This analysis is useful only in terms of defining experts who are clear 'outliers' and those who are part of the 'mainstream', and is therefore helpful in defining a 'consensus'. Thus, while there is significant variability among expert opinions, this particular subset of experts is the most consistent cluster. We also note (from Table 2) that the experts C, G and J span the three broad categories of experts, and thus we cannot say that any one of the three kinds of professional backgrounds of the experts is better than the others. It is also clear that any definition of consensus (i.e. fractional agreement) must be arbitrary.

Combined priors and posteriors

The analysis in the previous section has provided some useful hints about arriving at a reasonable approach to combining expert opinions. One approach is to argue that since the experts are in agreement for a majority of the situations, it makes sense to combine all expert opinions using the method described in Fig. 1. Figure 4a shows the prior obtained by combining all 11 experts for the smelter matte crushing area for metallic nickel. It is seen to be a roughly bimodal distribution with two experts predicting in the range 25–60%, and with the other nine predicting in the range 2–20%. The likelihood function is shown as a curve, and it peaks at the mean of the actual measurement (mean = 13.6%, SD = 8.6%). Figure 4b shows the corresponding posterior, and it very much resembles the likelihood function (i.e. the measured data). This phenomenon of the posterior resembling the measurement is seen in all the other cases (worksites and species) as well. A linear least-squares regression of the means of the priors versus the means of the measurements (for all worksites and species) had an R^2_{adj} of 0.73 and a slope of 0.52. However, a similar regression of the posteriors versus the measurements had an R^2_{adj} of 0.95 and a slope of 0.89. This is most likely the consequence of the combined priors having a rather broad distribution (reflecting the significant variability in expert opinions), while the sparse measurements have a relatively narrow distribution. This results in the measurements having a disproportionate influence on the posteriors.

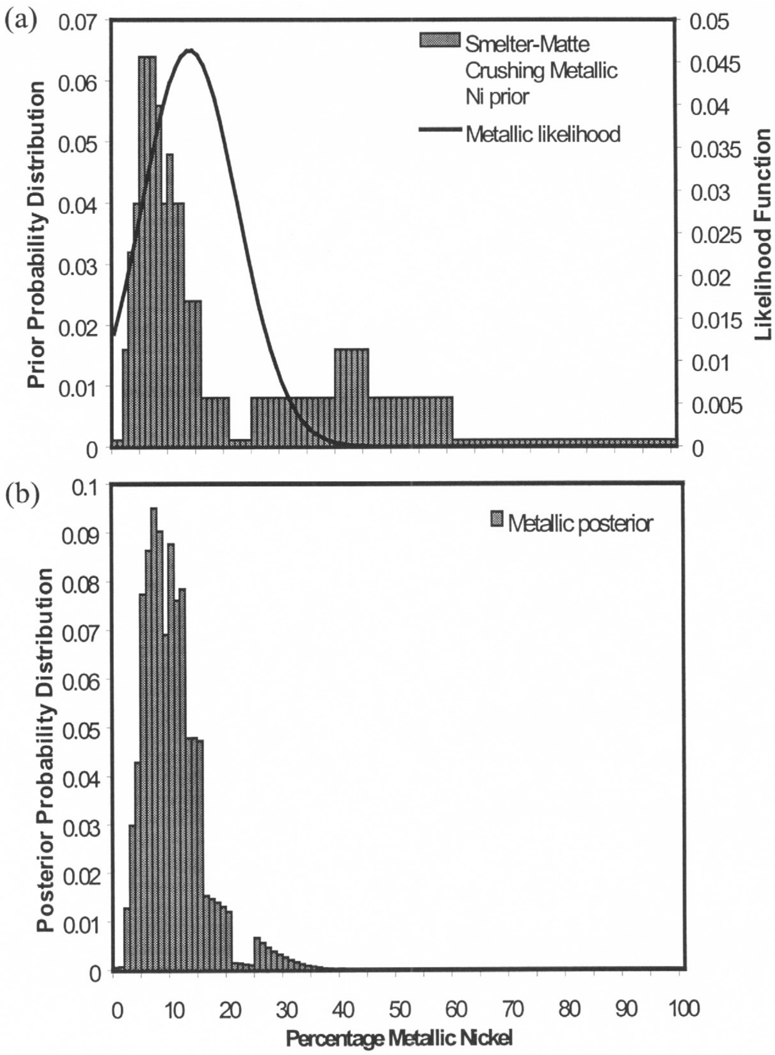

Figure 4. (a) Histogram showing the prior probability distribution obtained by combining all 11 expert priors. The smooth curve is the likelihood function based on actual measurements. (b) Histogram showing the combined posterior probability distribution for all 11 experts.

Another approach is to consider only the subset of experts who show the greatest level of consistency with each other and with the rest (i.e. experts C, G and J), and to combine only their priors and ignore the other experts. Figure 5a shows the prior obtained by combining all these three experts for the smelter matte crushing area for metallic nickel (same situation as Fig. 4a). The three experts agree among themselves and have a narrow distribution in the range 2–12%. As before, the likelihood function is shown as a curve. Figure 5b shows the corresponding posterior, and it quite closely resembles the prior function. This time, a linear regression of the means of the priors (combining the priors of only the three experts) versus the means of the measurements (for all worksites and species) had an R^2_{adj} of 0.62 and a slope of 0.59. A similar regression of the posteriors versus the measurements had an R^2_{adj} of 0.87 and a slope of 0.83. Thus, it appears that using only a subset of experts who are consistent among themselves increases the influence of the combined prior on the posterior. This is because combining the priors of all 11 experts resulted in a distribution that was much less precise than combining the priors of the consistent subset of experts.

We therefore chose to use this most consistent cluster of experts to obtain combined priors and then updated them to obtain combined posteriors. Table 6 shows the means and the standard deviations of the actual measurement data, the combined expert priors and the corresponding posteriors for all the four species for all 12 worksites. For most of the cases, we see that the posterior mean is between the mean of the data and the mean of the prior. With a few excep-

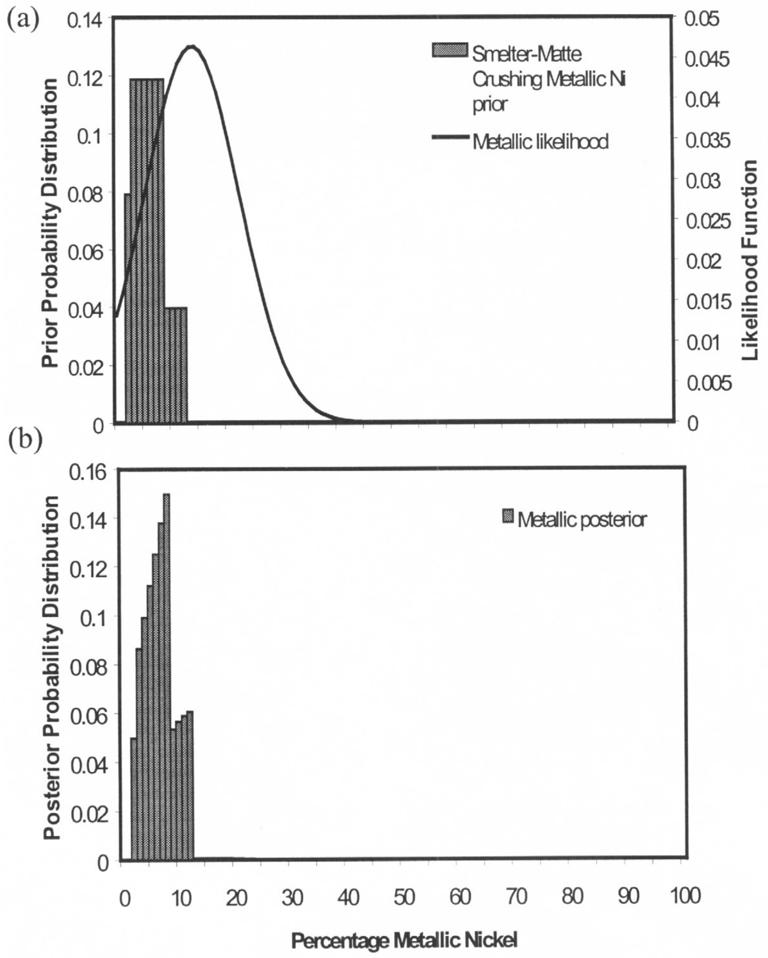

Figure 5. (a) Histogram showing the prior probability distribution obtained by combining priors of three experts who form the most consistent cluster. The smooth curve is the likelihood function based on actual measurements. (b) Histogram showing the posterior probability distribution for the experts comprising the most consistent cluster.

tions, the posterior is typically at least as precise or more precise than either the measurements or the priors. However, in some instances (e.g. soluble nickel in the mill-tipple), the prior is so different from the much more precise measurement data that the posterior mean is essentially the same as the measurement mean. Another interesting case is when the prior and measurement means are very similar even though the standard deviations of both are large (e.g. oxidic nickel in the smelter converter aisle). In such cases, the posterior mean is of course very similar to the prior and the measurement; however, the standard deviation is much reduced. That is, we obtain a much more precise estimate with the posterior.

CONCLUDING REMARKS

Many situations in occupational exposure assessment are marked by sparse data that are, by themselves, quite insufficient for reliably assessing risk. Professional hygienists have therefore supplemented these objective but scant data with subjective judgments that are claimed to be derived from their experience and knowledge. While this practice is widespread, there has been no systematic study of 'expert judgment' or the 'art' of occupational hygiene. Using a conceptual framework based on Bayes's theorem, new insights have been gained into the contribution that can be provided by carefully chosen experts.

This research, employing 11 experts who estimated the percentage of four nickel species in 12 worksites, provides a large dataset from which useful inferences can be drawn about the quality of expert judgments and the variability among the experts. A well-designed questionnaire that provided succinct information about the processes and baseline data served to calibrate the experts. The experts came from a variety of backgrounds, though all of them worked in occupational hygiene.

Table 6. Means and standard deviations of the actual measurements, combined expert priors (for the three experts C, G and J), and posterior distributions

Worksite	Nickel species	Measurement		Expert prior		Posterior	
		Mean	SD	Mean	SD	Mean	SD
Mill-tipple	Soluble	5.8	1.8	25.2	11.2	5.8	1.8
	Sulfidic	61.8	4.2	45.1	7.9	57.1	6.8
	Oxidic	24.3	2.3	29.2	10.6	25.0	2.1
	Metallic	8.1	3.6	6.0	12.9	5.2	2.4
Mill-grinding	Soluble	7.9	1.5	27.4	12.6	8.5	2.0
	Sulfidic	50.3	19.4	40.7	11.1	41.9	8.3
	Oxidic	34.1	18.7	32.4	9.5	31.6	5.3
	Metallic	7.7	5.6	5.5	12.7	3.6	2.0
Smelter-flash furnace	Soluble	3.8	0.5	15.9	12.1	4.6	0.6
	Sulfidic	69.2	12	45.2	10.4	52.3	8.3
	Oxidic	17.9	4.7	35.0	11.3	23.5	3.5
	Metallic	9.1	9.6	7.1	12.7	4.9	2.6
Smelter-converter aisle	Soluble	8	4.7	13.5	12.5	9.4	3.3
	Sulfidic	34.8	19.8	43.2	10.7	41.6	7.9
	Oxidic	43.6	19.9	41.6	10.2	41.3	7.2
	Metallic	13.5	5.6	39.5	10.1	15.6	7.6
Smelter-matte crushing	Soluble	4.3	2.8	13.8	12.7	6.2	2.1
	Sulfidic	73.3	12	45.0	11.9	57.5	9.9
	Oxidic	8.9	8.4	38.1	11.8	21.6	8.4
	Metallic	13.6	8.6	8.5	12.5	6.9	3.1
Smelter-matte process	Soluble	3.6	2.2	13.8	12.4	5.2	1.6
	Sulfidic	67.8	13.8	40.1	10.5	48.3	10.6
	Oxidic	16.5	10.1	44.2	10.6	31.7	9.2
	Metallic	12.2	6.5	8.1	12.6	6.7	2.6
Smelter-FBR	Soluble	4.3	1.3	11.5	12.6	4.6	1.2
	Sulfidic	24.4	16.3	37.2	11.9	33.3	8.2
	Oxidic	67.4	21.3	35.4	11.7	42.8	11.7
	Metallic	4	5.2	7.5	12.7	4.8	2.4
Smelter-Cottrell	Soluble	8.5	3.2	13.6	12.9	8.9	2.6
	Sulfidic	22	8.7	41.9	10.4	33.1	6.2
	Oxidic	64	2.2	43.7	11.7	61.8	2.8
	Metallic	5.5	3.3	7.0	12.8	4.9	2.3
Refinery-feed receiving	Soluble	2	3.1	11.8	12.2	5.5	1.8
	Sulfidic	10.1	7.7	34.3	14.3	15.7	5.6
	Oxidic	77.9	19.3	53.4	13.2	60.7	10.6
	Metallic	10	10.7	9.5	13.0	7.3	3.9
Refinery-TBRC	Soluble	6.7	0.8	9.4	12.8	6.7	0.8
	Sulfidic	9.2	5.9	16.2	13.2	10.8	4.3
	Oxidic	77.3	4.5	67.9	11.9	75.9	3.6
	Metallic	6.7	1.9	13.1	12.4	7.1	1.6
Refinery-ball mill	Soluble	2.6	3.7	8.1	12.5	4.6	2.2
	Sulfidic	14	11.6	13.7	13.4	11.6	5.7
	Oxidic	51.2	21.8	70.1	11.8	68.4	9.2
	Metallic	32.1	29.7	16.9	12.8	16.4	9.3
Refinery-packaging	Soluble	1.5	1.3	7.6	12.5	2.2	1.0
	Sulfidic	4.1	2.5	13.4	13.7	4.8	2.0
	Oxidic	62.7	10.6	72.1	11.9	69.9	6.1
	Metallic	31.9	11.7	15.7	12.6	17.6	6.7

There was a very high degree of agreement among the experts. A majority of the experts agreed among themselves 92% of the time, while almost two-thirds agreed 73% of the time. This was true irrespective of whether we used expert priors or posteriors to analyze the data. This, coupled with the fact that the experts came from varied backgrounds related to occupational hygiene, seems to suggest that there is indeed some broad body of specialized knowledge that the experts are drawing on to reach similar judgments. In other words, one could make the case that there is indeed such a thing as expert judgment in occupational hygiene.

One caveat is that the situation presented in this paper deals with normalized quantities (i.e. the percentages of nickel species that add up to 100%), where the domain is not large. However, in situations where the domain of the quantity to be estimated is large (e.g. exposure levels that can vary over several orders of magnitude), the performance of the experts may not be as good. In such situations, a more structured assessment where experts provide inputs to exposure models might be more useful (Cherrie and Schneider, 1999; Ramachandran, 2001).

This study also develops a matrix by which one can identify which experts tend to agree with each other most of the time (i.e. who represents the 'mainstream'), and thus can define a 'consensus' opinion. A cluster of experts who most consistently agreed with each other was identified. Interestingly, this cluster consisted of three experts—one from each category of expert. It seems that one type of expert is not necessarily any better than any other kind, and expertise does not necessarily require intimate familiarity with the worksite.

Combining all 11 expert priors which were then updated using Bayes's theorem resulted in posteriors that were very similar to the measurements. This is because the combined priors are much less precise than the measurements. Of course, using 11 experts for any exposure assessment exercise is quite rare in occupational hygiene practice where one or at most two experts are typically used. In this research, using a smaller set of three experts led to the priors having more influence on the posteriors.

The knowledge basis in this study is quite extensive and specific. It ranges from specialized knowledge such as the expected ordering of the magnitude of the different species to the more mundane fact that the fractional speciation should add up to 100%. Such a breadth and depth of knowledge is needed for useful elicitation of expert opinion. This suggests that in a more general exposure assessment exercise the experts must be chosen with care to ensure that they possess sufficient knowledge of the situation, and that the background anchoring information provided to the experts must be considerable so that they have an adequate knowledge basis for providing inputs.

Finally, the Bayesian framework has been used in this work to develop posterior means and standard deviations of the percentages of the four nickel species in the 12 worksites of interest in the company. As shown in Table 6, these represent estimates of the nickel speciation that are at least as precise as—and most of the time better than—those provided by the sparse measurement data from the Vincent *et al.* (2001) field study. This is because they incorporate additional historical information (from Vincent *et al.*, 1995) as well as the expert judgment of carefully selected occupational hygiene scientists. However, this should not be interpreted to imply that expert judgment can replace data. Rather the combination of expert judgment and sparse data is better than the sparse data alone. In this example, we have shown that the expert judgment exercise has indeed enhanced the quality of our knowledge of the exposure 'fingerprints' for the nickel industry worksites studied. For occupational hygiene exposure assessment, our experience suggests that such expert judgment methods can provide a cost-effective means by which information about worksite hazards can be improved and refined.

APPENDIX

Each expert has a prior belief, $P(\mu|a,b)$, for the mean percentage (μ) of a particular nickel species in the population. Here a and b are parameters of the uniform distribution—they specify an interval within which the percentage mean is supposed to lie. This prior belief is updated with the measured data. The posterior distribution of μ is computed from Bayes's theorem as,

$$P(\mu\,|\,a, b, \text{sample data}) = \frac{P(\mu\,|\,a, b)P(\text{sample data}\,|\,\mu)}{P(\text{sample data})} \quad \text{(A1)}$$

where $P(\text{sample data}|\mu)$ denotes the 'likelihood' of the data, in this case a normal distribution with mean μ and known standard deviation σ. It is easily shown that the posterior distribution $P(\mu|a,b,\text{ sample data})$ is a truncated normal distribution. This may be simulated simply by using

$$Z = \mu + \sigma^* Q(F[(a-\mu)/\sigma] + U^*\{F[(b-\mu)/\sigma] - F[(a-\mu)/\sigma)\}) \quad \text{(A2)}$$

in which U is a uniform random variable drawn from (0,1), Q is the standard normal quantile function and F is the standard normal distribution function. The truncated normal distribution is essentially the normal distribution restricted to an interval on the real line. Thus, while the usual normal distribution can take any value on the real line, the truncated distribution can only take values on the specified interval. In our setting, we chop off the two tails of the normal distribution, as specified by the lower and upper limits of the prior distribution. Strategies for simulating from truncated distributions, including the above, are discussed in Gelfand *et al.* (1992).

Acknowledgements—The experimental data central to the ideas in this paper were obtained during field research supported by funding from Inco Ltd and Falconbridge Ltd. We are grateful to Dr Bruce Conard of Inco Ltd for his encouragement and suggestions during the extension of that work described in this paper. We are also especially grateful to the 11 experts who participated in the study.

REFERENCES

ACGIH. (2001) Threshold limit values for chemical substances and physical agents, and biological exposure indices. Cincinnati, OH: American Conference of Governmental Industrial Hygienists.

Carlin BP, Louis TA. (1998) Bayes and empirical Bayes methods for data analysis. Boca Raton, FL: Chapman & Hall/CRC Press.

Cherrie JW, Schneider T. (1999) Validation of a new method for structured subjective assessment of past concentrations. Ann Occup Hyg; 43: 235–45.

Cooke RM. (1991) Experts in uncertainty: opinion and subjective probability in science. Oxford: Oxford University Press.

Doll R *et al.* (1990) Report of the international committee on nickel carcinogenesis in man (special issue). Scand J Work Environ Health; 16: 1–82.

Gelfand AE, Smith AFM, Lee TM. (1992) Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. J Am Stat Assoc; 87: 523–32.

Hawkins NC, Evans JS. (1989) Subjective estimation of toluene exposures: a calibration study of industrial hygienists. Appl Ind Hyg J; 4: 61–8.

Kromhout H, Oostendorp Y, Heederik D, Boleij JSM. (1987) Agreement between qualitative exposure estimates and quantitative exposure measurements. Am J Ind Med; 12: 551–62.

Little RJA, Rubin DB. (1987) Statistical analysis with missing data. New York: Wiley.

Makridakis S, Andersen A, Carbone R *et al*. (1984) The forecasting accuracy of major time series methods. Chichester: John Wiley.

Post W, Kromhout H, Heederik D, Noy D, Duijzentkunst RS. (1991) Semiquantitative estimates of exposure to methylene chloride and styrene: the influence of quantitative exposure data. Appl Occup Environ Hyg; 6: 197–204.

Ramachandran, G. (2001) Retrospective exposure assessment using Bayesian methods Ann Occup Hyg; 45: 651–67.

Vincent JH, Tsai PJ, Warner JS. (1995) Sampling of inhalable aerosol with special reference to speciation. Analyst; 120: 675–9.

Vincent JH, Ramachandran G, Kerr SM. (2001) Particle size and chemical species ‘fingerprinting’ of aerosols in primary nickel production industry workplaces. J Environ Monitor; 3: 565–74.

Zatka VJ, Warner JS, Maskery D. (1992) Chemical speciation of nickel in airborne dusts: analytical method and results of an interlaboratory test program. Environ Sci Technol; 26: 138–41.