ORIGINAL ARTICLE

Using a representative sample of workers for constructing the SUMEX French general population based job-exposure matrix

A Guéguen, M Goldberg, S Bonenfant, J C Martin

Occup Environ Med 2004;**61**:586–593. doi: 10.1136/oem.2003.010660

See end of article for authors' affiliations

Correspondence to:
Prof. M Goldberg,
INSERM Unité 88-IFR 69,
Hôpital National de
Saint-Maurice, 14, rue du
Val d'Osne, 94415
Saint-Maurice Cedex,
France; Marcel.Goldberg@st-maurice.inserm.fr

Accepted 29 January 2004

Background: Job-exposure matrices (JEMs) applicable to the general population are usually constructed by using only the expertise of specialists.
Aims: To construct a population based JEM for chemical agents from data based on a sample of French workers for surveillance purposes.
Methods: The SUMEX job-exposure matrix was constructed from data collected via a cross-sectional survey of a sample of French workers representative of the main economic sectors through the SUMER-94 survey: 1205 occupational physicians questioned 48 156 workers, and inventoried exposure to 102 chemicals. The companies' economic activities and the workers' occupations were coded according to the official French nomenclatures. A segmentation method was used to construct job groups that were homogeneous for exposure prevalence to chemical agents. The matrix was constructed in two stages: consolidation of occupations according to exposure prevalence; and establishment of exposure indices based on individual data from all the subjects in the sample.
Results: An agent specific matrix could be constructed for 80 of the chemicals. The quality of the classification obtained for each was variable: globally, the performance of the method was better for less specific and therefore more easy to assess agents, and for exposures specific to certain occupations.
Conclusions: Software has been developed to enable the SUMEX matrix to be used by occupational physicians and other prevention professionals responsible for surveillance of the health of the workforce in France.

Availability of exposure data is an essential part of any occupational health programme. Exposure data are needed to organise workplace monitoring and to control the occupational environment, to estimate the burden of occupational factors on the population's health, and are a critical component of epidemiological studies.[1] However, reliable exposure data are difficult to collect at the population scale and are rarely available, while occupational health surveillance would greatly benefit from tools aimed at giving valid estimates of the prevalence of exposure to chemicals according to occupations and economic sectors.

Among the various methods of exposure assessment, job-exposure matrices (JEMs) have become increasingly popular. Their general principle is based on the construction of a database that associates occupations (the rows of the matrix) with data about exposures to various hazards (the columns).[2] Linking the individual work history data of the subjects included in a case-control or cohort study with a JEM enables exposures to be attributed to the subjects. Despite some methodological limits, mainly due to their lack of specificity compared with individual assessment of exposure at the workplace,[3] JEMs have decisive advantages, because they can be used in very large scale surveys in which the traditional methods for assessing occupational exposures may be impossible to implement. Mainly intended for epidemiological purposes, especially for dealing with past exposures,[4] JEMs may also be used for routine surveillance of exposure at an individual or collective scale.[5]

Two principal approaches have been proposed for constructing JEMs.[6] Some are constructed a priori. They rely on the expertise of specialists who systematically review all the jobs and attribute to each indices that characterise its usual exposure to the toxic agents under study: intensity, frequency, and probability. This method is usually used for matrices applicable to the general population, when representative exposure measurements are not available for all the jobs likely to be encountered among the subjects.[4] Industrial cohort studies, on the other hand, sometimes have access to representative exposure measurements; JEMs can thus be constructed a posteriori, by regrouping the jobs to maximise the inter-group variance for exposure values and minimise the intra-group variance, and provide the best possible contrast between groups.[7 8]

The objective of this work was to construct a JEM for a long list of chemical agents from data collected from a large sample of French workers representative of the main economic sectors, that may be used by occupational health professionals for surveillance purposes. The construction of the SUMEX matrix was based on segmentation methods designed to optimise the grouping of jobs and thereby obtaining groups with homogeneous exposure prevalence.

METHODS

Data

The data available to us came from the cross-sectional SUMER-94 survey in France, carried out by occupational physicians on behalf of the Ministry of Labour.[9] Its principal objective was to describe the populations exposed to different types of occupational hazards, according to economic activity and company size as well as worker age, gender, and

Abbreviations: CART, classification and regression tree; JEM, job-exposure matrix; NAF, Nomenclature d'Activités Française; PBB, polybrominated biphenyl; PCB, polychlorinated biphenyl; PCS, Nomenclature des Professions et Catégories Socioprofessionnelles; TLV, threshold limit value

Main messages

- The SUMEX job-exposure matrix was constructed using a segmentation method, from data collected in a large sample of French workers representative of the main economic sectors.
- Software has been developed to use the SUMEX matrix easily through menus by querying by industry, occupations, or chemical agents.
- In spite of certain limits, the SUMEX matrix is a useful tool for occupational health professionals, providing reasonably valid information about exposures likely to occur either at the individual or at the workplace level.

Policy implications

- Exposure data according to occupations and economic sectors are needed for occupational health surveillance at a population level.
- Job-exposure matrices may be useful for routine surveillance of exposure to chemicals at an individual or collective scale.

occupation. The SUMER survey is planned to be regularly repeated; a new survey is currently in progress.

In France, contrary to most other countries, occupational physicians are in charge of monitoring workers' exposure. They are trained in occupational hygiene during their four years' specialised residency, and it is mandatory that they devote at least a third of their working time in such tasks, besides the annual medical visits that are also mandatory; there are about 7000 occupational physicians in France. Only employees of the private sector are monitored by occupational medicine: self-employed persons (including farmers and most of the agricultural workers), civil servants, public companies workers (such as employees of the national electricity or railroad companies), and some other specific categories do not benefit from this medical surveillance and were not included in the SUMER-94 survey. Finally, the survey covered a large part of the population of workers in France (about 12 millions out of a workforce of about 15 million).

A two stage sampling scheme was used: first, a sample of occupational physicians was established on the basis of their willingness to participate; then, all the physicians who volunteered were asked to randomly select a sample of the workforce that they monitored regularly, including both white and blue collar employees, following a sampling procedure designed by the principal investigators of the SUMER-94 survey. Finally 1205 occupational physicians working in different economic sectors each questioned on average 40 workers from June 1994 to June 1995; in all, 48 156 workers were included in the SUMER-94 sample. The participating physicians had to fill in a specific questionnaire describing the characteristics of the employing company of each worker, as well as the main characteristics of the workers themselves (age, gender, occupation, working schedule, type of contract). Regarding exposure, the questionnaire investigated 102 different chemicals that could possibly be present in the working environment (table 1), as well as some biological and physical agents and organisational factors. The survey was intended to cover only the last typical work week.

Based on measurements carried out in the workplace, completed by an interview with the workers about the tasks they actually performed and the materials they used during the last typical work week, and using their own knowledge of the working environment of the companies they control on a regular basis, the physicians recorded the presence or absence of the 102 chemical agents under investigation at the job during the last typical work week. For each chemical agent reported to be present, the physicians assessed the duration of exposure during that week in four categories (<2 hours; $\geqslant 2$ to <10 hours; $\geqslant 10$ to <20 hours; $\geqslant 20$ hours). The level of exposure was the intensity during exposing tasks (not the eight hour average intensity), with consideration of collective and individual protections in four categories: low (defined as slightly superior to that of the general population, or at the detection limit); medium (less than 50% of the regulatory threshold limit value (TLV)); high (around 50% of the TLV); and very high (may exceed the TLV, or the level of the population known to be most highly exposed). Occasional exposures were recorded only if they occurred during the last typical week of work. The routes of exposure were not recorded. Physicians were instructed to also take into account passive exposures from the environment of the workstation; however, this was not clearly stated in the questionnaire itself. Exposing tasks were not recorded.

The current version of the JEM concerns only the chemical exposures. Only a part of the SUMER-94 database was made available to us: the data used to construct SUMEX concerned all 48 156 workers participating in the investigation and included the principal economic activity of the company, the worker's occupation, the presence or absence of exposure to each of the 102 chemical agents during the last week of work, as well as the duration and intensity of any exposure.

The companies' economic activities were coded with the official French nomenclature (Nomenclature d'Activités Française, NAF),[10] which is organised in five hierarchical levels. The first level includes 17 main industrial classes; the second, third, fourth, and fifth levels include, respectively, 31, 60, 240, and 700 industrial groups. The workers' occupations were coded with the official French nomenclature (Nomenclature des Professions et Catégories Socioprofessionnelles, PCS),[11] which is organised in four hierarchical levels. The first level comprises eight items; the survey included workers in only five of these: craftsmen, tradesmen and shopkeepers, managers and professionals, intermediate occupations, and salaried white and blue collar workers. The second, third, and fourth levels include, respectively, 24, 42, and 455 items.

Statistical methods

The objective was to construct a matrix in which the rows would represent the jobs, and the columns the chemical agents. A job in the SUMER-94 survey was defined by a combination of a NAF economic activity code, and a PCS occupational category code. Because the vast majority of jobs in the sample involved only a few workers, the jobs had to be grouped to obtain reliable estimates. We used a segmentation method for this purpose. Segmentation is a general method that can be applied to general problems of regression, regression with censored data, or discrimination between two or more groups.[12]

We used the classification and regression trees (CART) method[13 14] for the job consolidation. This method constructs a tree by successive binary splits of the population on the basis of the exposed or non-exposed to a given chemical status of the workers. CART considers that no stopping rule could be relied on to discover the optimal tree, and uses an algorithm that contains three steps. First, after having randomly divided the whole sample into a base sample and a test sample, an overfitting tree is grown, using the base

Table 1 Prevalence of exposure to 102 chemical agents among the 48 156 workers in the SUMER-94 survey, and estimation of the percentage of exposed workers in France; number of rows for the classified chemicals and classification quality

Agents	Number of exposed workers in the sample	% of exposed workers in France*	Row number	Classification quality
1-3 butadiene	63	0.1%	8	0.035
Acid anhydrides†	105	0.2%	–	–
Acrylamide	54	0.1%	4	0.035
Acrylonitrile, methacrylonitrile	26	<0.1%	2	0.001
Aldehydes	1430	2.6%	8	0.250
Alkyl ethers	166	0.3%	3	0.015
Allergenic plants	345	1.0%	11	0.072
Aniline, diethyl and dimethyl aniline	23	<0.1%	2	0.001
Animal dusts	202	0.5%	4	0.106
Antimony	38	<0.1%	4	0.007
Aromatic amines	149	0.3%	6	0.017
Aromatic halogenated and/or nitrated hydrocarbons†	327	0.6%	–	–
Arsenic	38	<0.1%	2	0.000
Asbestos	415	0.8%	11	0.172
Azides†	42	0.1%	–	–
Barium, soluble compounds	68	0.1%	4	0.040
Benzene	281	0.6%	9	0.056
Beryllium and compounds	28	<0.1%	5	0.119
Bis-chloromethyl ether†	8	<0.1%	–	–
Borons and compounds	68	0.1%	3	0.044
Cadmium	55	0.1%	4	0.171
Carbon monoxide	564	1.3%	3	0.103
Carbon sulphide†	14	<0.1%	–	–
Cement	1462	2.6%	22	0.444
Chromium VI, chromic acid and compounds†	262	0.4%	–	–
Coal and petroleum based tar, pitch, and asphalt	335	0.5%	5	0.086
Cobalt and compounds	120	0.2%	7	0.134
Combustion smoke	843	1.6%	5	0.033
Cresols†	56	0.2%	–	–
Crystalline silica	479	0.8%	7	0.111
Cyanoacrylates	184	0.3%	7	0.023
Dimethyl formamide	90	0.2%	4	0.015
Diverse solvents: acetates, esters, and ketones	1937	3.2%	25	0.071
Enzymes	119	0.2%	3	0.009
Epichlorhydrine	49	0.1%	2	0.001
Epoxy resins	622	1.1%	10	0.035
Ethylene glycol and low molecular weight polymers†	340	0.6%	–	–
Ethylene oxide†	45	0.1%	–	–
Fluorosilicates	28	<0.1%	4	0.001
Furfural and furfurylic alcohol	28	<0.1%	4	0.035
Halogenated solvents	2465	4.2%	25	0.068
Halogens	473	0.8%	6	0.081
Hexane	185	0.4%	12	0.023
Hydrazine and compounds†	65	0.1%	–	–
Hydrides	75	0.1%	5	0.065
Hydrofluoric acid†	244	0.4%	–	–
Hydrogen cyanide	98	0.2%	4	0.008
Hydrogen phosphide	10	<0.1%	–	–
Iron oxides	284	0.5%	7	0.033
Isocyanates and prepolymers	552	1.0%	7	0.088
Lead and compounds	611	1.0%	10	0.053
Manganese and compounds	54	0.1%	6	0.084
Mercury	95	0.2%	6	0.073
Metal carbonyls†	61	0.1%	–	–
Methanol	416	0.8%	9	0.039
Methyl halides†	68	0.2%	–	–
Methyl pyrrolidone	39	<0.1%	3	0.003
Mineral dusts	1334	2.1%	16	0.078
Mineral oils	2662	4.4%	36	0.217
Monoalkylated ethers of ethylene glycol	586	1.1%	8	0.014
Mycotoxines†	17	<0.1%	–	–
Nickel and compounds	207	0.4%	6	0.070
Nitrogen oxides	163	0.3%	5	0.038
Nitrosamines	61	0.1%	3	0.053
Organophosphorous insecticides	301	1.2%	4	0.175
Other oils	1177	2.1%	9	0.071
Other oxidants, peroxides	261	0.6%	4	0.031
Other resins	769	1.3%	6	0.026
Ozone†	99	0.2%	–	–
Paraquat and diquat	132	0.5%	3	0.099
PCB and PBB†	15	<0.1%	–	–
Persulphates	241	0.7%	3	0.448
Petroleum based hydrocarbons	1608	3.2%	16	0.159
Petroleum based solvents	2683	4.8%	19	0.154
Phenols, halogenated and nitrated compounds	399	0.9%	3	0.141
Phosgene and other carbon oxyhalides†	29	<0.1%	–	–
Phosphorus and its salts	51	<0.1%	3	0.002
Plant dust	625	1.3%	7	0.087
Polycyclic aromatic hydrocarbons†	355	0.6%	–	–

Table 1 Continued

Agents	Number of exposed workers in the sample	% of exposed workers in France*	Row number	Classification quality
Selenium and compounds†	11	<0.1%	–	–
Sintered metallic carbides	89	0.1%	5	0.154
Solvents: alcohols other than methanol	2091	3.8%	22	0.082
Strong acids	2304	4.5%	16	0.085
Strong bases	2863	5.9%	17	0.138
Styrene	236	0.4%	5	0.022
Sulphates and alkylated sulphides	33	<0.1%	4	0.003
Sulphites	74	0.2%	7	0.015
Sulphur oxide	89	0.2%	5	0.009
Surfactants	2757	6.0%	19	0.196
Synthetic mineral fibres (glass and ceramic)	329	0.5%	6	0.024
Tetrahydrofuran	110	0.2%	3	0.072
Thallium†	6	<0.1%	–	–
Tin, inorganic salts	88	0.2%	3	0.003
Vanadium and compounds†	8	<0.1%	–	–
Vinyl chloride monomer	58	0.1%	3	0.014
Volatile acrylates	141	0.2%	4	0.010
Volatile aliphatic amines	449	0.7%	5	0.035
Volatile methacrylates	158	0.3%	4	0.065
Volatile nitriles	41	<0.1%	2	0.002
Vulcanisation fumes	148	0.2%	3	0.164
Welding fumes	1647	3.0%	21	0.210
Wood dust	937	1.5%	24	0.311

*Private sector only; total workforce of about 12 million in France (SUMER data;[9] see text).
†Could not be classified.

sample; second, this overfitted tree is successively pruned back so that a sequence of nested sub-trees is obtained; finally, the test sample is run through all the sub-trees of the sequence in order to select the optimal tree—that is, the one having the lowest cost estimated by the test sample. As the aim is to obtain good probability estimates of exposure and not only good classification decisions, class probability trees were grown instead of classical classification trees. The tree cost is defined as the sum of the terminal nodes costs; for the class probability tree, cost at each node is the mean square error between a dummy variable indicator for exposed/not exposed and the observed within node probability of exposed. The quality of a tree, defined as [1 − relative cost] ranges from 0 to 1 and is comparable to the R^2 that measures the explanatory value of multiple regression or the inter-group variance ratio divided by the total variance obtained in an analysis of variance. When the quality of the sub-trees of the sequence decreased since the first split, that means that the data are non-informative for the considered chemical agent, and the CART algorithm created no tree.

The complete matrix was constructed in two stages: (1) grouping of occupations according to exposure prevalence (percentage of exposed workers) by using the CART method; and (2) establishment of exposure indices (probability, duration, and intensity) within the terminal nodes.

Table 2 Distribution of jobs observed in the SUMER-94 survey

Number of workers per job	Number of jobs	Percentage of jobs	Number of workers	Percentage of workers
1	7759	57.5	7759	16.1
2–4	3880	28.7	10023	20.8
5–9	1066	7.9	6851	14.2
10–29	599	4.4	9172	19.1
30–99	150	1.1	6912	14.3
100 or more	43	0.3	7439	15.4
Total	13497	100.0	48156	100.0

CONSTRUCTION OF THE SUMEX MATRIX

Descriptive results of the SUMER-94 survey

Table 1 shows the prevalence of exposure to the 102 chemicals among the 48 156 workers in the SUMER-94 survey, and the estimated proportion of exposed workers in France among the target population (about 12 million workers). Only five agents (strong bases, mineral oils, halogenated solvents, petroleum based solvents, and surfactants) had an exposure prevalence greater than 5%, and nine agents (aniline, diethyl and dimethyl aniline, bis-chloromethyl ether, mycotoxines, polychlorinated biphenyls (PCBs) and polybrominated biphenyls (PBBs), hydrogen phosphide, carbon sulphide, selenium and compounds, and thallium and vanadium and compounds) had an exposure prevalence of

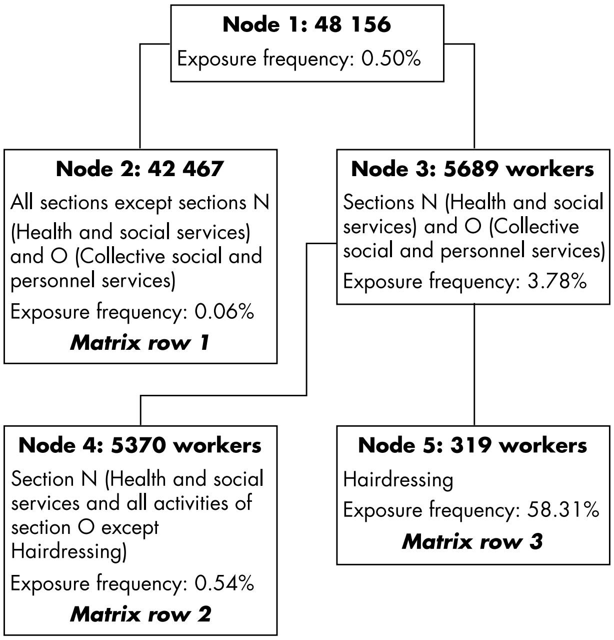
Node 1: 48 156
Exposure frequency: 0.50%
Node 2: 42 467
All sections except sections N (Health and social services) and O (Collective social and personnel services)
Exposure frequency: 0.06%
Matrix row 1
Node 3: 5689 workers
Sections N (Health and social services) and O (Collective social and personnel services)
Exposure frequency: 3.78%
Node 4: 5370 workers
Section N (Health and social services and all activities of section O except Hairdressing)
Exposure frequency: 0.54%
Matrix row 2
Node 5: 319 workers
Hairdressing
Exposure frequency: 58.31%
Matrix row 3

Figure 1 Persulphate specific tree.

less than 0.05%, corresponding to fewer than 24 exposed workers among the 48 156 in the sample.

The sample counted a total of 671 different NAF codes, 420 different PCS codes, and a total of 13 497 different jobs. The theoretical number of jobs resulting from the combination of 700 NAF codes and 434 PCS codes being approximately 300 000, the sample thus did not include most of the possible combinations. Among the non-represented jobs, some corresponded to improbable combinations and others to plausible but very rare combinations. Moreover, 86% of the sample jobs involved four workers or fewer (table 2).

Construction of the matrix rows

The CART algorithm has been applied to each of the 102 chemical agents. The whole sample has been randomly divided into a base sample (n = 24 129) and a test sample (n = 24 027). For 80 chemical agents, an optimal tree was selected and its terminal nodes constituted the matrix rows. This method yields a specific matrix for each chemical agent. Each matrix was composed of a variable number of rows, depending on the number of terminal nodes obtained. It made up a ''column'' of the final SUMEX matrix. SUMEX is thus made up of a set of matrices, each specific for a chemical agent. For 22 chemical agents, CART created no tree (table 1). All these agents were used very rarely, or the exposure is most often present only in the working environment and not generated by the tasks directly performed by the workers, such as for polycyclic aromatic hydrocarbons, suggesting that occupational physicians did not properly assess exposures not directly generated by the workers' tasks.

Exposure indices

For each row of the agent specific matrices, three exposure indices were computed: probability (that is, prevalence among all 48 156 workers in the sample), and average duration and intensity of exposure among exposed workers.

Example: persulphate exposure

The results for exposure to persulphates are presented to illustrate the matrix construction.

Definition of the rows

The prevalence of persulphate exposure was 0.50% in the total sample (0.44% in the base sample and 0.56% in the test sample). The optimal tree selected by the CART algorithm is represented in fig 1, where the percentages of workers in each of the rows of the persulphate matrix and the exposure prevalence by row are reported. The quality of the tree estimated by the base and test samples are respectively 0.450 and 0.447; these values were close, and the quality of the tree estimated by the test sample was as good as the one yielded by the base sample, indicating that the tree was not overfitted.

The optimal tree splits twice and distributes the population into three classes. Results are given for the whole sample. The first split was performed on the 48 156 workers. This was a split of the NAF code and created nodes 2 and 3. Node 2 contains the 42 467 workers in all industries, except those with a NAF code of N (health and social services) and O (collective social and personnel services); the mean exposure prevalence for node 2 was 0.06%. Node 3 contains 5689 workers with a NAF code of N or O, with a mean exposure prevalence of 3.78%.

Table 3 Prevalence of persulphate exposure in nodes 1–5 of the tree

		Exposure prevalence	
Matrix row	Node	Base sample	Test sample
	1	0.44%	0.56%
1	2	0.03%	0.09%
	3	3.56%	3.99%
2	4	0.53%	0.55%
3	5	56.29%	60.12%

Table 4 Distribution of the mean exposure duration and intensity per week for each row of the persulphate specific matrix among the exposed workers (percentages)

	Row 1	Row 2	Row 3
Duration			
Less than 2 hours	64.0	28.6	29.5
From 2 to fewer than 10 hours	20.0	42.9	49.7
From 10 to fewer than 20 hours	8.0	25.0	14.2
20 hours or more	8.0	3.6	6.6
Total	100.0	100.0	100.0
Intensity			
Low	43.5	30.8	27.6
Medium	39.1	19.2	55.3
High	17.4	50.0	14.7
Very high	0.0	0.0	2.4
Total	100.0	100.0	100.0

Node 2 is not split. The second split concerned workers in node 3 and created nodes 4 and 5. Node 5 contains the 319 workers whose NAF code corresponds to the activity ''hairdressing'': the mean prevalence of persulphate exposure was 58.3%. Node 4 contains the 5370 workers in section N or O, except those involved in ''hairdressing''; the mean prevalence of persulphate exposure was 0.54%.

Table 3 reports the prevalence of persulphate exposure in nodes 1–5 for the base sample and the test sample. These exposure prevalences are close, showing that there was no overfitting.

Exposure duration and intensity

The duration and intensity of exposure for each line of the persulphate matrix were calculated from the data of all the exposed workers in the sample. Table 4 reports the distributions by category of exposure duration and intensity for the workers exposed in each row. Some of the jobs included in rows 1 and 2, where exposure is rare (less than 1% of the workforce in the industrial sectors which were grouped in these rows), may nevertheless be heavily exposed to persulphate, such as specific occupations in the perfume industry, or in chemistry research.

Other agents

Table 1 summarises the results for the other 79 agents for which we obtained a satisfactory job consolidation. It reports the number of rows (that is, the number of terminal nodes of the segmentation tree) and the quality of the classification obtained for each agent specific matrix in the total sample. The quality of the trees was highly variable, ranging from values over 0.3 for persulphate, cement, and wood dust, to values under 0.001, as for arsenic (0.0004) or epichlorhydrine (0.0007). Globally, the performance of the method seemed better for less specific and therefore more easy to assess agents. It was also usually better when the exposure is concentrated into few occupations, by contrast with agents that are largely disseminated over many jobs.

Use of the SUMEX matrix

Software has been developed to enable the SUMEX matrix to be used by occupational physicians and other health and safety professionals responsible for the surveillance of the health of the workforce in France.[15] It can be used easily through menus that facilitate access to the database by querying by industry, occupations, or chemical agents. As most of the occupational physicians are in charge of the surveillance of workers of numerous small companies, they usually use the matrix by asking about the hazards occurring in specific industrial sectors or jobs; the software gives the list of all the chemicals that exist in the industrial sector or job, ordered by decreasing prevalence, and for each of them the exposure indices (probability, intensity, duration), as well as the number of exposed subjects and the total number workers of the industry or occupation interviewed in the SUMER-94 survey.

When querying by chemical agents, the SUMEX software gives an ordered list of all industrial branches and jobs where the chemical is present, with the associated exposure indices; this type of query may be useful when planning measurement and control surveys.

DISCUSSION

The SUMER-94 survey gives a picture of the prevalence of exposures to different types of occupational hazards in France. Detailed results describing the prevalence of exposure according to age, gender, socioeconomic status, size of the company, industry, and occupation were published by the Ministry of Labour.[9] However, the very large number of different occupations and industries represented in the sample (671 different industrial sector codes, 420 different occupation codes, yielding 13 497 different combinations, 86% of them involving four workers or fewer) did not allow publishing reliable estimates for specific occupations and industries.

In order to group the jobs in a coherent way regarding specific exposures, we used a segmentation method to construct a job-exposure matrix from the data gathered in this large sample of 48 156 workers representative of most industries in France, in which their regular occupational physicians assessed their exposure to 102 chemicals. We had no control on the SUMER-94 survey, conducted by the Ministry of Labour, which made a part of the database available to us.

The validity of the matrix relies mainly on the representativeness of the sample and on the quality of the exposure data collected by the occupational physicians. Compared to the 1990 French population census, the representativeness of the sample for the private sector (covering approximately 12 million workers out of a total workforce of 15 million), was found to be very close to the population of workers in France for the principal demographic and occupational characteristics.[9]

Substantial inter-physician variability was observed for the prevalence of exposure in the same occupation. This variability may reflect real differences in intra-job exposure due to the possible variation in the working conditions for a single occupation, or differences in inter-physician judgement. Without proper reference data, we cannot assess the respective roles of these two sources of variability. This problem may influence the validity of the job groupings in the SUMEX matrix because of the method we used to construct it. To assess the extent of this potential bias, we studied the influence of each of the 1205 occupational physicians for each of the 80 chemicals. We calculated the exposure prevalence obtained for the job groups without the data from each physician. For 62 chemicals, the effect of the most influential physician modified the exposure prevalence by less than 10%; for the other 18, the estimate of exposure prevalence within job groups was more dependent on a single physician. This may be explained by either a real ''physician effect'' or by the fact that the physician was one of the few who surveyed workers in a given occupation; the latter is true for the rare occupations. In all, the exposure prevalence of only 18 of 627 job groups constructed for the 80 chemical was likely to have been influenced by a real physician effect.

The quality of the exposure data collected by the occupational physicians survey was based primarily on the physicians' knowledge of the occupational environment of the workers they monitor supplemented by personal interviews with the individual workers. Two points need to be considered: the existence of a specific exposure, and the frequency and intensity of exposure for the chemicals that were present at the workstation. As a side study of the SUMER-94 survey, a small sub-sample of 144 workers was randomly selected and interviewed again by experienced occupational hygienists about the same work week within a few days after the interview with the occupational physician; the second interview was based on an open questionnaire derived from the method first established by Gérin and colleagues[16] and used in France for several epidemiological studies.[17] [18] The exposures were then coded and compared to the exposure data recorded by the occupational physicians. Globally, occupational physicians coded fewer exposures than industrial hygienists did. They identified 333 exposures occurring at the workstation versus 555 (73.1%); among them, only 15 were specific chemicals present in complex mixtures (such as engine exhaust or welding fumes), whereas the hygienists coded 120 such exposures. In addition, industrial hygienists also coded 190 exposures in the environment of the workstations, the occupational physicians having recorded only the exposures due the tasks performed by the workers, due to the wording of the questionnaire, which was not explicit about environmental exposure. Finally, when considering only broad categories of chemicals at the workstation, there were only a few differences between both sources regarding the presence of an exposure.[19] For exposures in the environment of the workstations and for specific chemicals present in complex mixtures, the SUMER-94 data do not seem reliable, and obviously underestimated the prevalence of exposures to agents such as manganese, iron oxides, hexavalent chromium and nickel from welding fumes, or polycyclic aromatic hydrocarbons, nitrogen oxides, and carbon monoxide from engine exhaust. Agricultural exposures also showed very low prevalences, due to the fact that in France most of the agricultural workers are self-employed and were not included in the survey.

Regarding the frequency of exposure, the data were based on interviews of the workers about their last typical work week, and there is no reason to think that under or overestimation occurred. For intensity, the level of exposure was established on the basis of measurements at the workstation when available; these measurements may have been performed in different circumstances, the vast majority being for the purpose of regulatory exposure limits control. As intensity exposure was considered only for exposing tasks, while it was categorised in reference to TLV, this could have resulted in an overestimation of the average exposure intensity in the SUMER-94 survey; in order to improve the current version of SUMEX we are planning to compare the intensity estimates with the French COLCHIC database[24] whenever possible. However, the matrix being intended to help the occupational physicians by warnings about potential problems at the workplace, inaccuracies in intensity estimates is not a serious concern for the day to day occupational heath surveillance.

The quality of the classification obtained varied according to the chemical agents (table 1). It was satisfactory for some and mediocre for others. The quality of classification reflects the homogeneity of the groupings of jobs according to exposure prevalence; it is best when all exposed jobs are gathered in one single row, and deteriorate when exposed jobs are scattered among different rows. As it is not likely that occupational physicians recorded exposures that do not exist in specific jobs, the main reason for a bad quality of the classification for some chemicals is that the NAF and PCS nomenclatures were designed for economic statistics and not constructed according to exposure to occupational hazards; exposure to some widely used substances occurs among trades which are disseminated through very different industries. The CART segmentation method was used to optimise the grouping of jobs regarding the prevalence of exposure, but could not solve problems that are induced by inadequate nomenclatures. This is a limitation of any JEM constructed on the basis of official nomenclatures. In fact, table 1 shows that when exposure is concentrated among a small number of occupations where most of the workers are exposed, the quality of the tree was good, but it deteriorated when exposure involves many occupations for which only some workers are actually exposed. The comparison of the exposure data collected by the occupational physicians and by experienced industrial hygienists showed also that in general, the occupational physicians were less specific in their assessment of exposure, but that the reproducibility between the two methods can be considered satisfactory for large categories of chemicals. This is well reflected by the classification process, the quality of the trees being usually better for less specific agents. Finally, it is probable that the structure of the matrix reflects with a good degree of validity the average distribution of the prevalence of exposure at the workstation among jobs for large categories of chemicals, or when exposure is specific to certain occupations. For more specific chemicals or when exposures are scattered among numerous occupations with only a small proportion of workers actually exposed, the exposures may be less contrasted between the job groupings; moreover, it is very likely that SUMEX does not reflect the exposure to chemicals present in the environment of the workers, but not generated by the tasks they perform. The overall bad quality of the groupings would have been a real problem if one wanted to use SUMEX to estimate the prevalence of exposure to chemicals in specific jobs. However, we think that this is not a real concern regarding the purpose of the matrix and the way the software works; it is intended to warn the user about potential exposures, and answers only queries about specific occupations and industries, without any global statistical feature.

Finally, a group of six experienced occupational physicians who did not participate in the SUMER-94 survey were asked to examine the job groups obtained and the values of the exposure indices; they also used the matrix for several weeks in their own regular practice with the workers they had to examine during that period. This was not a regular evaluation of the accuracy and usefulness of the matrix, which would imply a heavy protocol, but an informal exercise. Globally, the occupational physicians found that the matrix was reasonably realistic. They also found SUMEX useful for their practice: the usual way of using the matrix was to query the database each time they felt that they were not fully informed about the environment of the workplace of some workers, or to prepare their planned site visits. Finally, the occupational physicians recommended that SUMEX should be largely distributed.

The CART segmentation method[13 14] is appropriate to optimise the job groupings according to exposure prevalence. The SUMEX matrix therefore uses all of the information available from a large representative sample. Such an approach is comparable to the a posteriori methods of matrix construction, which rely on the analysis of representative exposure measurements among workers. The a posteriori approach has been used until now only in specific industries, where it was possible to carry out measurement surveys for some chemicals.[7 8 20 21] In several countries, databases pool the results of exposure measurements for various agents in numerous industries.[22–24] Nonetheless, no country has available sufficiently representative measurements of common occupational exposures to numerous chemical agents for the thousands of jobs that can be encountered in the general population to construct JEMs based solely on statistical analysis of such databases.[25 26]

Existing matrices for the general population have therefore been constructed a priori by relying on experts and using databases as a complement.[27–29] The method used here is an effective compromise between the a priori and a posteriori approaches, and it preserves the practical advantages of using the substantial quantity of available demographic, occupational, and economic data that use the French national classifications.

SUMEX has some limitations in its current version. As is most often the case for general population based JEMs, the validity of the matrix could not be formally assessed with proper reference data. The quality of the job groupings varies largely regarding chemicals, and therefore SUMEX should be used with caution, as acknowledged by the software we developed. In spite of the problems linked to the nature of the data collected within the SUMER-94 survey and the limited size of the sample, we believe that the SUMEX matrix is a useful tool for occupational health professionals, able to provide them with reasonably valid information regarding exposures likely to occur either at the individual or at the workplace level.

ACKNOWLEDGEMENTS

The authors thank Doctor N Sandret and Mesdames O Héran-Leroy and N Yahou, who coordinated the SUMER-94 survey, as well as the Department of Research, Surveys, and Statistics (Direction de l'animation de la recherche, des études et des statistiques), the Labour Medical Inspection (Inspection médicale du travail), and the Department of Labor Relations (Direction des relations du travail) of the Ministry of Labor and Solidarity (Ministère de l'Emploi et de la Solidarité), who provided access to the data. The SUMEX project received funding from the French Ministère de l'Emploi et de la Solidarité–Direction de l'animation de la recherche, des études et des statistiques.

Authors' affiliations

A Guéguen, M Goldberg, S Bonenfant, J C Martin, INSERM Unité 88-IFR 69, Hôpital National de Saint-Maurice, 14, rue du Val d'Osne, 94415 Saint-Maurice Cedex, France

REFERENCES

1 **Armstrong BK**, Saracci R, White E. *Principles of exposure measurements in epidemiology.* Oxford, New York: Oxford University Press, 1994.

2 **Hoar S**. Job exposure matrix methodology. *J Toxicol Clin Toxicol* 1983–84;**21**:9–26.

3 **Plato N**, Steineck G. Methodology and utility of a job-exposure matrix. *Am J Ind Med* 1993;**23**:491–502.

4 **Hémon D**, Bouyer J, Berrino F, *et al.* Retrospective evaluation of occupational exposures in cancer epidemiology: a European concerted action of research. *Appl Occup Environ Hyg* 1991;**6**:541–6.

5 **Goldberg M**, Imbernon E. The use of job exposure matrices for cancer epidemiology research and surveillance. *Arch Public Health* 2002;**60**:173–85.

6 **Goldberg M**, Hémon D. Occupational epidemiology and assessment of exposure. *Int J Epidemiol* 1993;**22**(suppl 2):S5–9.

7 **Goldberg M**, Kromhout H, Guénel P. Job exposure matrices in industry. *Int J Epidemiol* 1993;**22**(suppl 2):S10–15.

8 **Guénel P**, Nicolau Molina J, Imbernon E, *et al.* Design of a job-exposure matrix on electric and magnetic fields: selection of an efficient job classification for workers in thermoelectric power production plants. *Int J Epidemiol* 1993;**22**(suppl 2):S16–21.

9 **Ministère de l'emploi et de la solidarité (France)**. *Expositions aux contraintes et nuisances dans le travail. SUMER 1994. Les dossiers de la DARES.* Paris: Ministère, 1999.

10 **Institut National de la Statistique et des Études Économiques (France)**. *Nomenclature d'Activités et de Produits (NAF/CPF)*. Paris: INSEE, 1992.

11 **Institut National de la Statistique et des Études Économiques (France)**. *Nomenclature des Professions et Catégories Socioprofessionnelles (PCS)*. Paris: INSEE, 1994.

12 **Zhang H**, Crowley J, Sox HC, *et al*. Tree-structured statistical methods. In: Armitage P, Colton T, eds. *Encyclopedia of biostatistics*. Chichester, 1998;**6**:4561–73.

13 **Breiman L**, Friedman JH, Olshen RA, *et al*. *Classification and regression trees*. Belmont, NY: Wadsworth, 1984.

14 **Salford Systems**. CART Software, version 2.5. San Diego, CA: Salford Systems, 1991–1997.

15 **Guéguen A**, Martin JC, Bonenfant S, *et al*. *Matrice emplois-expositions SUMEX* (CD-ROM). Paris: Éditions INSERM, 2000.

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

17 **Hours M**, Dananché B, Févotte J, *et al*. Bladder cancer and occupational exposures. *Scand J Work Environ Health* 1994;**20**:322–30.

18 **Cordier S**, Bergeret A, Goujard J, *et al*. Congenital malformations and maternal occupational exposure to glycol ethers. *Epidemiology* 1997;**8**:355–63.

19 **Févotte J**, Danaché B, Cachon M, *et al*. Un autre regard sur l'enquête SUMER-94: l'évaluation des expositions professionnelles de salariés par jugement d'expert [in French]. *Documents pour le Médecin du Travail* 1997;**70**:147–53.

20 **Kromhout H**, Symanski E, Rappaport SM. A comprehensive evaluation of within- and between-worker components of occupational exposure to chemical agents. *Ann Occup Hyg* 1993;**37**:253–70.

21 **Morgan D**. Occupational exposure databases and their application for the next millenium: symposium framework and workshop introduction. *Appl Occup Environ Hyg* 2001;**16**:111–14.

22 **Boiano JM**, Delon Hull R. Development of a national occupational exposure survey and database associated with NIOSH hazard surveillance initiatives. *Appl Occup Environ Hyg* 2001;**16**:128–34.

23 **Kauppinen T**. Finnish occupational exposure databases. *Appl Occup Environ Hyg* 2001;**16**:154–8.

24 **Vincent R**, Jeandel B. COLCHIC—Occupational exposure to chemical agents database: current content and development perspectives. *Appl Occup Environ Hyg* 2001;**16**:115–21.

25 **Brederode D**, Linker F, Marquart H, *et al*. Recording of data of individual measurements of occupational exposure: guideline of the Dutch society of occupational hygiene. *Appl Occup Environ Hyg* 2001;**16**:122–7.

26 **Cherrie J**, Sewell C, Ritchie P, *et al*. Retrospective collection of exposure data from industry: results of a feasibility study in the United Kingdom. *Appl Occup Environ Hyg* 2001;**16**:144–8.

27 **Kauppinen T**, Toikkanen J, Pukkala E. From cross tabulations to multipurpose exposure information systems: a new job exposure matrix. *Am J Ind Med* 1998;**33**:409–17.

28 **Pannett B**, Coggon D, Acheson, eds. A job-exposure matrix for use in population-based studies in England and Wales. *Br J Ind Med* 1985;**42**:777–83.

29 **Ferrario F**, Continenza D, Piani P, *et al*. Description of a job-exposure matrix for sixteen agents which are or may be related to respiratory cancer. In: Hogstedt C, Reuterwall C, eds. *Progress in occupational epidemiology*. Amsterdam: Excerpta Medica, 1988:379–82.