



Exposure assessment in industry specific retrospective occupational epidemiology studies.

N S Seixas and H Checkoway

Occup. Environ. Med. 1995;52;625-633
doi:10.1136/oem.52.10.625

Updated information and services can be found at:

<http://oem.bmj.com/cgi/content/abstract/52/10/625>

These include:

References

9 online articles that cite this article can be accessed at:

<http://oem.bmj.com/cgi/content/abstract/52/10/625#otherarticles>

Rapid responses

You can respond to this article at:

<http://oem.bmj.com/cgi/eletter-submit/52/10/625>

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article

Notes

To order reprints of this article go to:

<http://journals.bmj.com/cgi/reprintform>

To subscribe to *Occupational and Environmental Medicine* go to:

<http://journals.bmj.com/subscriptions/>

Exposure assessment in industry specific retrospective occupational epidemiology studies

Noah S Seixas, Harvey Checkoway

Abstract

Quantitative estimation of exposure for occupational epidemiology studies has received increasing attention in recent years and, as a result, a body of methodological literature has begun to take form. This paper reviews the generic issues in the methodology of exposure assessment, particularly methods for quantitative retrospective assessment studies. A simple framework, termed an exposure data matrix (EDM), for defining and analysing exposure data is proposed and discussed in terms of the definition of matrix dimensions and scales. Several methods for estimation, interpolation, and extrapolation, ranging from subjective ratings to quantitative statistical modelling are presented and discussed. The various approaches to exposure assessment based on the EDM concept are illustrated with studies of lung disease among coal miners and other dust and chemically induced chronic occupational diseases. The advantages of validated statistical models are emphasised. The importance of analysis and control of errors in exposure assessments, and integration of the exposure assessment and exposure-response processes, especially for emerging occupational health issues, is emphasised.

(*Occup Environ Med* 1995;52:625-633)

Keywords: exposure assessment; exposure data matrix; retrospective studies

Quantitative estimation of exposure has become a central focus in occupational epidemiology over the past decade as a result of the increasing emphasis put on exposure-response characterisation for occupational hazards. Through the many studies conducted explicitly to determine optimal methods of estimating exposure, a body of methodological literature has begun to emerge. Several comprehensive reviews on data sources and methods of assessment of exposure are available.¹⁻⁵ In this article we attempt to synthesise some of the concepts represented in this literature into a somewhat generalised approach for plant or industry specific retrospective exposure assessments. General goals of exposure assessment are considered in terms of the errors associated with quantitative assessments, and a

methodological framework for organising and analysing exposure information is proposed. Examples are included of how such a framework may be used, as drawn from the scientific literature, with emphasis on retrospective studies of chemicals and dusts.

Concepts of exposure and dose

Determination of the effect an agent has on health, whether in toxicology or epidemiology, requires that the dose of the agent to a person be defined as accurately as possible. Dose may be defined as "the amount of a substance that remains at the biological target during some specified time interval".³ Thus, dose is characterised in four dimensions: *identity* (a substance), *form* (at the target site, and implying bioavailability), *concentration* (amount), and *time* (specified time interval).^{6,7} In a toxicology experiment, these dimensions of dose may be relatively well defined in terms of specifying the agent used, the amount or concentration delivered, the route of exposure, and the duration and frequency with which the agent is delivered.⁸ In an epidemiological context, in which we observe rather than deliver the dose, defining these dimensions of dose to each person, and especially to a population, poses a substantial challenge.

The process of exposure assessment is thus one of translating a measurable quantity of exposure to an approximate, or estimated dose received by each study subject. This representation of dose for epidemiological analysis is referred to as an *exposure metric*. Error is inevitably introduced in the process of estimating an exposure metric from some observed or measured set of exposure data or information. It is important to understand that here "error" does not imply mistakes, but rather, as in a mathematical expression of the relation between the true dose of a substance received by a person, D_i , and the observed exposure information, Z_i ,

$$D_i = Z_i + a + \varepsilon_i.$$

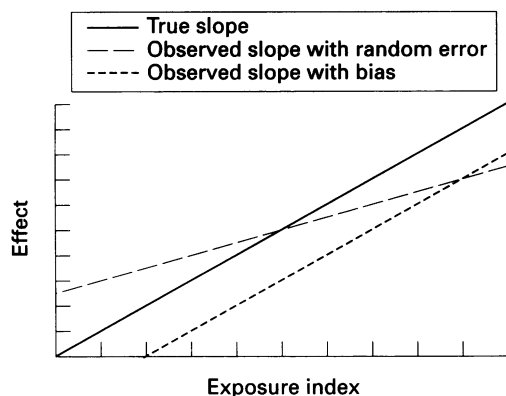
In this expression, the error includes both a fixed or systematic bias, a , and a random error component, ε_i . The figure shows the effects of these two types of error in the context of a simple linear regression. Note that a systematic error, or change in intercept caused by a non-zero a , will alter the predicted effect at a given level of exposure. A random error in exposure,

Department of
Environmental Health,
University of
Washington, Box
357234, Seattle,
WA 98195-7234, USA
N S Seixas
H Checkoway

Correspondence to:
Dr N S Seixas, Department
of Environmental Health,
University of Washington,
Box 357234, Seattle,
WA 98195-7234, USA.

Accepted 7 July 1995

Effects of biased and imprecise exposure estimates on a hypothetical exposure-response relation.



however, stemming from a large variance of ε_p , results in an attenuation of the observed exposure-response function. Attenuation of an exposure-response relation due to random error or misclassification of exposure, whether it be in the context of linear regression, logistic regression, or more complex models has been explored extensively in recent years.⁹⁻¹² The degree to which an exposure assessment is successful in providing accurate estimates of the true dose-response function, is therefore dependent on the degree to which both the systematic and random aspects of error can be controlled. That is, *the goal of exposure assessment is the minimisation (or characterisation) of the systematic and random errors in estimates of individual dose derived from available exposure information, across the study population.*

Systematic and random error may be introduced in an exposure assessment in the process of defining any of the four dimensions of exposure: identity, form, concentration, and time. In fact, much of traditional industrial hygiene practice is aimed at methods to minimise these errors (see⁶ for further definition and discussion of these dimensions). These aspects are frequently not under the control of the epidemiological investigator because the data collection typically has occurred before a study is conducted. Moreover, industrial hygiene measurements are usually made for purposes other than epidemiological research, such as compliance monitoring.¹³

Exposure assessors must draw on industrial hygiene research to recognise, quantify, and account for the errors associated with exposure information. The primary concern for most retrospective epidemiological studies is the paucity of useful information. The problem of missing data or information from historical periods, often necessitates extrapolation of recent information to earlier periods. These early periods may have special significance because of the relatively high exposures that occurred during early years of a process. Furthermore, there may be little or no data available on certain categories of subjects, such as those in low exposure jobs. Varying amounts and quality of data among plants in a study of many sites poses yet another challenge. For instance, in an industry wide study, one plant may have done much less monitoring than others, or certain areas of high exposure may have received all of the industrial hygiene effort, leaving the lower exposed areas unmonitored. Exposure information may also be insufficiently specific. For example, work

history records frequently list job classifications that are non-specific with respect to tasks, and hence to actual exposures.¹⁴ In a related manner, if exposure data are collected on a sample of workers, they may inadequately represent exposures to others.¹⁵ Thus, estimating exposure for times and places for which insufficient data are available is fraught with uncertainty and may introduce a significant degree of error into the exposure assessment.

Given these common limitations of the available information on exposure, estimation of exposure levels is particularly problematic in the face of the high degree of variability typically seen in occupational environments. Variability in exposure concentrations from day to day can range over orders of magnitude. For example, geometric SDs as high as 4.5 have been commonly noted.^{15,16} Also, even within groups of subjects chosen to have similar exposure distributions, geometric SDs between people are commonly as high as 2.0, representing over 10-fold differences between the 97.5th and 2.5th percentile of individual means.^{15,17}

A method to assess exposure

Table 1 shows the generic steps in a retrospective assessment of exposure. The first steps of gathering, describing, and validating available information have, as already mentioned, been discussed extensively elsewhere^{18,19} and will not be reviewed in detail here. Defining a structure in which the available information may be analysed, and specifying an approach for using that information are the components of exposure assessment that are emphasised. Development of an appropriate exposure metric and exposure-response modelling, which is the final step of the research, has been discussed elsewhere^{20,21} and will not be considered explicitly.

A simple structure for organising and understanding information for exposure assessments is an exposure data matrix (EDM). By organising information in an EDM, the assessment process, with each of its component steps of coding, describing, analysing, estimating, and linking exposure data to study subjects, can be defined more explicitly. Furthermore, the EDM structure can be used as an analytical framework for almost any exposure assessment process. The first step in the development of an EDM is to define its dimensions and scales. Figure 2 shows a generic two dimensional EDM. Two dimensions are depicted, but an EDM may be defined with only one dimension, for instance, exposure zone, or with as many dimensions as the available information can usefully support.

In figure 2, the two dimensions are labelled time and job as examples of dimensions that are

Table 1 Steps in assessment of epidemiological exposure

1	Gather and characterise all relevant exposure-related information
2	Evaluate data for errors including bias and precision
3	Define exposure data matrix for linkage to study subjects
4	Estimate exposure levels
5	Summarise exposure metrics for each subject
6	Estimate exposure-response relations

commonly used. Other dimensions might include department, plant, or geographic region, type of machine or process, presence of exposure controls, etc. For studies of mixed exposures, a dimension to represent each different agent might be used and for some studies in which different methods of analysis are used for the same agent during different historical periods, methods of analysis could occupy a dimension of the EDM. Similarly, different types of information, each providing important aspects of the exposure profile, such as personal and area sampling data or airborne and biological monitoring data, may be available for the study, and form dimensions of the EDM. The decision of how many, and which dimensions should be included for a particular study depend primarily on the extent of the data available and a judgment of the importance of each factor in representing exposure.

The next step is to define the scale of each of the EDM dimensions. The scale may be continuous, or in ordered categories. For instance, a time axis could be continuous, with all other dimensions referred to a specific point in time. More commonly, time would take the form of ordered categories such as months, years, or longer periods. A categorical scale is most appropriate for characteristics such as jobs that have no natural ordering. The use of categories also implies that exposures within a category are homogeneous. In fact, within the context of the EDM, the concept of a homogeneous exposure zone^{22 23} is most easily identified as an individual cell of the matrix, irrespective of how the dimensions are defined.

A primary part of defining matrix scales is determining the degree to which the available information should be grouped. For instance, one might have sufficient information on individual study subjects to include a dimension for subject, with each person represented as a separate category of that dimension. More commonly, subjects are identified by their association with a grouping factor such as job title. With job as a grouping factor, it is assumed that the factor is predictive of exposure. Such predictive factors must also be readily linked to each subject through their work history. Frequently, many jobs are further grouped under the assumption that they have indistinguishable exposures, or because there is insufficient detail in any exposure data, or work history information, to allow them to be kept separate. The specificity of the exposure assessment is reduced as categories are grouped to

form larger groups with decreasingly homogeneous exposures. The extent of grouping is generally determined on the basis of judgment about the greatest degree of specificity that the data meaningfully support. Analysis of the errors associated with different degrees of grouping, and their implications for exposure-response relations, are beginning to show quantitative solutions to the grouping problem²⁴, (and personal communication from Heederik and Attfield).

After the dimensions and scales of the matrix are defined, the available data are distributed into the relevant cells. Although the EDM is most easily understood for quantitative exposure measurements, each cell could also contain qualitative data. For instance, each cell could contain pertinent information such as production data or subjective exposure estimates made by plant personnel. If industrial hygiene data are available, then placing them into this format allows for calculation of arithmetic means, or other parameters of the exposure distribution that might be of interest. Biological monitoring data may also be used in this manner.

Under most circumstances, a relatively simple matrix (even if multidimensional) may be defined. In certain situations useful data may be available that do not fit this form. For instance, task specific exposure information may be available and task oriented exposure assessment may actually be a very efficient means by which exposure can be measured.^{26 27} Under some circumstances, task may represent an additional dimension of the EDM. For instance generic tasks such as machine operating, maintenance, or cleaning that can be defined for multiple jobs, may be called uniform job categories²⁸ and be included in the matrix. In other circumstances, the tasks may be specific to a single job category, with each job having a unique set of component tasks. In this case, one might think of task as a characteristic nested within another dimension of the matrix.

Exposure in each cell of the matrix may be estimated once the pertinent exposure data are defined in the EDM structure. With the available exposure data defined in the matrix, the missing data for subjects, groups (such as jobs or departments), or time periods are readily observed and estimation methods including imputation, extrapolation, and a variety of modelling approaches can be applied. These methods are discussed later. Finally, the estimates of exposure in each cell of the matrix may be used to compute several exposure metrics for each subject that are then used in the exposure-response analysis.

Thus, an EDM may be as a single dimension, for instance, job category with one type of quantitative exposure data represented, or may be a complex multidimensional matrix including both crossed and nested factors with several types of exposure information. Although, ideally one would want many highly specific dimensions including all bits of available information, the matrix definition ultimately rests on the type and amount of data available, and the ability to relate the information in a meaningful manner to the study subjects' work experience.

		Dimension 1: eg time period					
		T ₁	T ₂	T ₃	T ₄	• • •	T _T
Dimension 2: eg job	J ₁	Y _{1,1}	Y _{2,1}	Y _{3,1}	Y _{4,1}	• • •	Y _{T,1}
	J ₂	Y _{1,2}	Y _{2,2}	Y _{3,2}	Y _{4,2}	• • •	Y _{T,2}
	J ₃	Y _{1,3}	Y _{2,3}	Y _{3,3}	Y _{4,3}	• • •	Y _{T,3}
	J ₄	Y _{1,4}	Y _{2,4}	Y _{3,4}	Y _{4,4}	• • •	Y _{T,4}
		•	•	•	•	• • •	•
J _J	Y _{1,J}	Y _{2,J}	Y _{3,J}	Y _{4,J}	• • •	Y _{T,J}	

Figure 2 Generic exposure data matrix

By structuring the data in an EDM, the relations between data components may be clarified, the available methods of analysis become more apparent, and the implications of particular methods in terms of error structures may be better understood.

Simple example of an EDM: respirable coal mine dust

A relatively straight forward example of the EDM approach is provided by the estimation of respirable coal mine dust exposures for a cohort of underground miners studied by the United States National study of coal workers' pneumoconiosis.^{29 30} The initial exposure assessment included only miners who worked in 36 mines in the period 1970–87 during which many thousands of personal respirable dust samples were collected by mine operators and government inspectors for compliance with the Coal Mine Safety and Health Act. The work history records collected for the study defined each miners' activity by the mine in which he or she worked and the start and end dates for each specific job held.

The first steps of the exposure assessment involved selecting and describing the relevant exposure data. Data from 314 000 samples collected at the 36 mines during the study period were used in the exposure assessment. The data quality was examined in detail, and some adjustments were made to account for identified biases.²⁹ In describing the data, it was found that, despite the size of the dataset, it was incomplete. Most of the data were collected within the first eight years, and 70% of the samples were associated with only 10 of about 150 jobs found in the miners' work histories. Furthermore, the cohort included miners working before these data were available; thus, exposure estimates for years before 1970 were also required to estimate cumulative exposure. Also, although the matrix was developed for use on this cohort from the study mines, the estimates derived could be extrapolated to other mines or miners. Thus, exposure levels could be directly calculated for some aspects of the miners' histories, but had to be estimated for portions of the matrix that had little or no data, and extrapolated to earlier periods or to mines other than those included in the original study.

A three dimensional EDM was defined by time, job category, and mine for the initial project that covered only the period during which data were available for the study mines.³⁰ Each dimension was scaled with relatively finely divided categories; time was categorised into single years (18 categories), job was classified with the 150 individual job codes recorded in

the work history and on the sampling data forms, and each of the 36 mines was classified separately. As a result, the three dimensional EDM with a total of almost 97 000 potential matrix cells was defined and the air sampling data were distributed into the cells. Arithmetic means (SEMs) were calculated in each cell (only about 10 000 cells contained data) and were linked to the miners' work histories. The cell means were used to calculate cumulative exposures for each miner as the primary chronic exposure metric. Because of the large number of job classifications, they were also grouped into four classes: face jobs involving work at the coal face, non-face jobs including maintenance, repair, and transportation jobs away from face operations, supervisory jobs, which are highly variable in nature, and surface jobs. Table 2 shows the number of strata and their average dust levels within these classes. No job specific means were calculated for the surface classification because of the limited sample size available for these jobs. Additional analyses of this matrix, and its extension to earlier periods and other mines are discussed later.

Methods of estimation

For an EDM with missing information in portions of the matrix, including earlier or future periods of time, various methods have been developed to estimate the needed information.

SUBJECTIVE ESTIMATION

In the relatively common situation in which there are few, or no data for exposure assessment, use of subjective evaluation about the presence, or level of exposure may be the only methods available. Such evaluations may be conducted by researchers, professional staff—for example, occupational hygienists—from the site or long term employees with knowledge of the specific processes and time periods being assessed. Inherent to the process is that the EDM used for subjective assessment initially should be defined in relatively few dimensions and with relatively crude categories. Typically, subjective ratings are provided on an ordinal scale, rather than as direct quantitative estimates of exposure intensity.

Several investigations have considered the validity of subjective estimation. Because they have primarily relied on quantitative data for validation, the estimates have referred mainly to current known conditions rather than historical conditions.^{31 32} These studies have shown that subjective raters are able to rank exposure levels with some validity, but that there is a high degree of variability between raters. Although jobs within a plant may be correctly ranked, the rankings across separate processes may be less accurate. In one investigation of toluene exposure, descriptive information about the process such as might be available for a historical reconstruction, led to inaccurate exposure estimates.³³ Provision of a limited number of quantitative results from personal samples can substantially improve the subjective estimation process.³⁴ Many questions such as the relative merits of individual assessments *v* a group consensus process, and the level of measure-

Table 2 Specific coal dust exposure estimates for job, mine, and year by broad job category³⁰

	Job category		
	Face	Non-face	Supervisory
Number of cells	4266	3903	1954
Mean samples/cell	50.5	17.3	5.1
Mean of mean concentration (mg/m ³)	1.46	1.05	0.58
Mean SEM concentration (mg/m ³)	0.63	0.38	0.23

ments—for example, number of categories, or quantitative *v* categorical estimation—that can be obtained with a subjective process, have not been fully answered.

SIMPLE ALGORITHMS AND MARGINAL MEANS

Interpolation and extrapolation of exposure data with a simple algorithm may be easily understood from a matrix. For instance, in a quantitative exposure-response analysis of leukaemia and benzene, a very limited set of measurements of benzene exposure was available for a variety of jobs in the rubber hydrochloride film production process for certain years.³⁵ Jobs were combined into a limited number of job categories (that is, the dimension of the matrix defined as jobs was defined with a limited number of categories) for which some data could be assigned. The algorithm used a stepwise set of estimates for exposure. Firstly, the mean within the matrix cell was used where data were available for a specific job category and year. Secondly, for a job category with some data gaps in certain years, simple linear interpolation was used to estimate concentrations for years with no data. Finally, for years in which no data were available before a certain year, the first point was extrapolated back in time. Forward extrapolation was done similarly for jobs in which data were non-existent after a certain year.

An alternative algorithm was developed for use in the coal dust assessment already described.³⁰ As described, a three dimensional matrix of mine, job, and year was defined and mean exposure values calculated in each cell for which data were available. These directly calculated means for job/mine/year cells accounted for 37% of the individual cells found in the cohort's work history. Again, a stepwise algorithm was adopted for estimating the exposures for matrix cells with missing data. Firstly, assuming that mines were closely related, the job/year means were calculated—that is, the marginal means across the mine dimension of the matrix. These marginal means accounted for an additional 30% of the miners' exposures. Next, the job dimension was reduced into the four broad job categories already discussed and means were calculated for the three dimensional matrix defined now by mine, year, and broad job category. These estimates accounted for an additional 10% of the work histories. Finally, data were further combined across the mine dimension by taking the job category and year means, which accounted for the remaining 23% of the miners' jobs. Thus, a stepwise algorithm that used marginal means across various dimensions of the matrix were used to estimate exposure where data were missing.

A variety of simple algorithms may be developed given the particular nature of the data available to exploit the most specific information available and to extrapolate the available data to periods with little or no information. In the benzene example,³⁵ the data were quite limited and the matrix was defined in a relatively simple fashion. Benzene exposure data for each job class were considered independently and extrapolation was conducted only in the time dimension. In the coal dust example in which

many more specific data were available, extrapolation was conducted across the mine and job dimensions.³⁰

EXTRAPOLATION WITH MULTIPLIERS

An alternative method of extrapolating a limited set of exposure data to times or places (or conditions) where fewer data are available is by development of factors, or multipliers, to account for those data. The need for this approach is quite common because one may frequently have available a set of exposure data for recent periods, or a new set of data are collected for the purposes of the study, but little or no data are available for earlier times. Extrapolation of the recent data to earlier periods, in which processes and conditions may have been substantially different, is an important problem. A wide variety of information may be used to develop factors for conducting this retrospective extrapolation.

Data-derived multipliers

In the coal dust example, it was necessary to estimate exposures for miners before the beginning of the dust sampling programme in 1970. The only systematic set of data available for this period was a survey conducted by the Bureau of Mines (BOM) in 1968–9. Because conditions were thought not to have changed substantially up to that time, these data were used to provide estimates for the earlier periods. The BOM data, however, were not available for all jobs and hence could not be used directly. The BOM data were instead used to develop a multiplier with which the more complete Mine Safety and Health Administration (MSHA) data after 1970, could be extrapolated to earlier periods on a job specific basis.³⁶ All jobs for which there were more than 10 samples in the BOM data were used and the average ratio of the BOM data to the MSHA data over the period 1970–2 was taken. This ratio (2.3) was then multiplied by the job specific MSHA data to obtain job specific estimates for the exposures before 1970. Use of the multiplier assumed that the reduction in exposures over this period was relatively uniform across jobs (both for those used to calculate the ratio and all other jobs) and mines, and that the concentrations estimated for the 1968–9 period were valid for earlier times, as well. In this case the matrix had job and time dimensions. Time was represented by two categories, the BOM survey period of 1968–9 (and all earlier time), and the MSHA data period of 1970–2 (subsequent time was represented by the estimates made directly from the MSHA data as already described). Thus, the average relation between these two time periods for selected jobs was used as the basis of extrapolation for all other jobs.

Subjective factors

Exposure levels determined for recent periods can be extrapolated back in time with subjective estimates of the significance of specific engineering changes occurring within a plant or department. For instance, in estimating historical exposures to formaldehyde, subjective exposure reduction factors associated with the

introduction of particular control technologies were applied to current exposures to estimate historical levels.³⁷

Subjective multipliers may also be used when current exposures as well are only estimated subjectively. For example, in a study of lung cancer associated with exposure to crystalline silica in the production of diatomaceous earth, the available exposure data were not directly used in forming exposure estimates.³⁸ Instead, jobs were assigned an exposure intensity score (none, low, moderate, high) on the basis of process observation and interviews with plant personnel. To extrapolate these estimates back in time, five historical periods were defined (before 1944, 1945–53, 1954–63, 1964–73, and 1974–87) and exposure multipliers for each period were established (12, 6, 2, 1.5, and 1, respectively). These factors were based on subjective assessment of the probable increases in concentration in each job for the historical periods. The calculated exposure metric was a stronger predictor of mortality from non-malignant respiratory disease than was years of dust exposure; however, the associations with mortality from lung cancer were equally strong for the two exposure indices.

Deterministic factors

Physical attributes of the process, exposure, or controls can be used to predict personal exposures. For instance, in a reconstruction of historical exposures to man made mineral fibres, the effects of changes in products, process, and controls on exposure levels were investigated in laboratory studies.³⁹ Experimentation was conducted to determine the importance to exposure levels of fibre dimension, addition of oil in the production process, changes in ventilation, and production rate and factors were derived that were associated with each of these process modifications. These factors were then applied to concentrations measured in the processes during a recent period, to estimate exposures historically from a detailed history of engineering and process changes at the plant.

One particularly novel approach to estimating past exposures that might be used directly, or as multipliers for current exposure levels is to recreate the actual historical process and measure exposures directly. Such experimentation has been conducted for man made mineral fibres (MMMF) during the production of rock-wool,⁴⁰ and for silica in granite sheds.⁴¹ This approach is dependent on the ability to recreate accurately the conditions that contribute most to exposure, and this requirement may always remain an elusive goal.

STATISTICAL MODELLING

Perhaps the most comprehensive approach to exposure estimation, and one that is most easily related to the EDM, is statistical modelling. This approach is powerful for several reasons. Firstly, information about the influence of each dimension of the matrix on exposure estimation may be determined through the modelling procedure. Secondly, although simple algorithms or deterministic factors may only draw on portions of the data matrix, such as a single job or the earliest data available, a statistical model

can draw on all available information to estimate particular exposures. Thus, a statistical model has the power to consider the matrix as a whole and uses the interrelation between the data segments to provide estimates of any one part of the matrix. Thirdly, models can be developed that weight the data in comparison to their relative importance. For instance, by using sample size cells with a large quantity of specific data may be weighted more heavily than cells with limited information. Similarly, a weighting factor could be provided for each cell based on the relative certainty or validity that the information was thought to represent.

There are also some inherent limitations to mathematical modelling—most importantly that the information derived from the model is dependent on the degree to which the model actually fits the data. Given that there may be times and places in which exposure conditions were unusual, and the information concerning those conditions were either unmeasured or not included in the model, the model predictions may contain a significant degree of error. Thus, when a statistical modelling approach is adopted, validation of the model (as described below) is of utmost importance.

The common approach is a simple linear model such as:

$$Y_{ji} = a + \sum \beta_j \mathcal{J}_j + \sum \beta_i \mathcal{T}_i + \varepsilon_i$$

This model represents the data that might be found in the matrix defined in fig 2, in which there are dimensions for job (\mathcal{J}) and time (\mathcal{T}) that are divided into categories, indicated by \mathcal{J}_j and \mathcal{T}_i respectively. Because these two dimensions are categorical, \mathcal{J} and \mathcal{T} represent indicator variables, and β_j and β_i are the estimated coefficients for each category. The individual exposure data Y_{ji} are distributed among the job and time categories and contain residual variability represented by ε_i . Introducing interaction terms into such a model ($\beta_{ji} \mathcal{J}_j \mathcal{T}_i$ for each combination of time and job) allows for more flexible relations between the different levels of each dimension.

One constraint on modelling is the number of categories that may be used. Estimation of the coefficients by least squares requires that there are at least as many data points as there are categories, and unless there are at least several data points for each parameter estimated, the parameter estimates will be very unstable. In practice, there may be a limit to the number of categories that can be estimated. For instance, in the coal mine matrix example, a statistical model would include 202 parameters (35 mine indicators, 17 year indicators, 149 job indicators, plus one intercept). Although the number of data points available would have allowed this, such a model would be highly unwieldy. Furthermore, if interactions among the three dimensions were considered, the number of parameters to estimate would be enormous. In this example, use of a simpler approach, the stepwise use of marginal means was adopted. An alternative approach of rescaling the dimensions into a smaller number of categories and with a statistical model for estimation may have proved equally useful. As noted earlier, combining categories involves

loss of specificity that the data might otherwise support.

Linear models may be used for exposure estimation in two similar, but related contexts. Firstly, the model may be used for interpolation of data in cells with little or no data. For instance, average concentrations of asbestos were estimated in a textile plant in which almost 6000 measurements were available over the years 1930–75.⁴² The matrix was defined with nine categories of zone (departments), four task categories, presence or absence of specific controls in each zone, and four time intervals. The linear model was estimated separately for each of the zones. The resulting regression coefficients were then used to provide estimates of mean exposure in each category of job and time. As a result of this modelling procedure, estimates of exposure were made for jobs and times where no data were available.

A second application of exposure modelling is extrapolation of exposure estimates for places or times when no specific quantitative data are available. An illustrative example is a retrospective study of exposure to and mortality from machining fluid particulates among car machinists.⁴³ For this study, only 394 samples were available for the period 1958–87 and most of them (243) were obtained in one plant. Before conducting the final analysis, linear models were used to define the matrix by considering the importance of time periods, plant, and sample type (area *v* personal). On the basis of these analyses, plant was combined into two categories, time was divided into three periods, and area and personal samples were not distinguished from one another. The final analytical matrix was thus defined as plant (two categories), time (three historical periods), fluid type (three types: straight oils, soluble oils, and synthetics) and operations (broadly grouped as grinding, machining, and assembly operations). The exposure estimates predicted were derived from a model that included each of these variables, plus the interaction between time and plant, and time and operation. As well as the use of estimated exposures for periods and plants contained within this dataset, the parameter estimates were also used to extrapolate to other plants and times based on the type of fluid in use and the operation (machining or grinding).

Another important application of statistical modelling is estimation of exposure to one agent from data on another related, or surrogate agent. In some cases, sampling may be conducted for a particular agent only because it is correlated with the suspected aetiological agent and its measurement is less expensive, or more feasible.⁴⁴ For instance, in a study of bladder cancer among aluminium smelter workers, benzene soluble matter in particulates was measured as a surrogate for benzo-*a*-pyrene.⁴⁵ The two measured entities formed two dimensions of the EDM, their relation could be discerned, and the results ultimately expressed in terms of exposure to benzo-*a*-pyrene.

Similarly, in a study of acute effects of machining fluid on respiratory outcomes, a simple measurement of thoracic particulates was used as an inexpensive personal measure-

ment technique. A suspected aetiological agent was bacteria, or some agent closely associated with bacteria (such as endotoxin), but these analyses were very time consuming and expensive. A subset of the air samples was therefore analysed for bacterial count and a linear statistical model, controlling for other aspects of the matrix such as department, time, shift, and fluid bacterial concentration, was developed and used for estimating personal bacterial exposures.⁴⁶

Statistical modelling of exposures has also been developed in a somewhat different context—in identifying the predictors, or determinants of exposure. Several investigators have used statistical models to identify exposure determinants for predicting exposures for epidemiological studies,^{47–49} and others have developed predictive models for the purposes of describing exposure data, and identifying potentially effective control strategies.^{50–51} For instance, a list of 23 variables were considered as potential factors in describing—or predicting—exposure to ethylene oxide in spice sterilisation.⁴⁷ By eliminating non-significant factors, the number of predictive variables was reduced to seven and these were again used to predict exposures for an epidemiological study of cancer mortality. In another study, alternative strategies to control machining fluids in the car industry were identified by building a linear model to describe the potential influence of ventilation characteristics, machine type, and fluid type.⁵⁰ In each instance, the model allowed consideration of the potential effects of altering one variable, while controlling for the influence of other variables in the system.

By understanding the alternative uses of linear modelling of exposure data and their common roots in the exposure data matrix, the meaning of exposure assessment becomes clearer. In effect, the definition of the dimensions and scales of the data matrix, either by initial judgments, or through statistical testing with an explicit model, is the identification and measurement of characteristics that determine exposure. To describe exposure, observable characteristics that have predictive value for exposure must be defined. The use of any particular factor, and its scale, is dependent on its ability to predict exposure to individual study subjects. Factors with little predictive value have little use in the assessment and can generally be dropped from the analysis.

Validation of the model

The use of linear statistical models raises the concern of the suitability, or validity of the predicted exposures. It is also useful to recognise that some form of modelling is involved in every approach to exposure estimation, including subjective assignment, simple algorithms, or complex mathematical modelling. Any of these approaches should ideally be subject to validation.

Validation methods for predictive statistical models have been reviewed.⁵² Generally, the available data are randomly split into model development and model validation subsets. The development subset is used to identify the

appropriate model—for example, the dimensions and scaling of the matrix—or possibly a set of competing models with alternative structures. The models are then used to predict a set of exposure estimates and the estimates are compared with the model validation subset. The comparison may take account of the bias (average difference between the predicted and observed exposures), precision (SD of the differences), or both, which is termed accuracy, or mean squared error (square root of the sum of the precision and the squared bias). Several alternative models may be constructed and the bias and precision of each compared to identify the best model. To provide the most accurate predicted values, the parameters of the selected model may be re-estimated with the combined (development and validation) dataset.

An alternative approach to validate exposure assessment is one in which the predictive value of an assessment for the risk of a well known outcome is used to validate the exposure metric for another related, but less known disease. An example is a study of risks related to silica in which the predictive value of estimated exposure to silica dusts for silicosis was used to validate the silica exposure metric, which was then used to examine its effect on risk of lung cancer.⁵³ Exposure validation of this sort is only useful in the limited context in which a single agent is related to two outcomes, and at least one is well characterised. Furthermore, there needs to be existing information on the shape of the exposure-response curve for the indicator disease to enable validation of the exposure assessment with respect to another disease. In the silica example, a linear, or at least monotonic relation between silica and silicosis could be assumed based on previous research.

Finally, validation of exposure models may be conducted by testing the exposure estimates derived from alternative models for their predictive value in the exposure-response model.^{20 54} This approach rests on the assumptions that there is an underlying exposure-response relation, and that the exposure metric that provides the highest effect estimate, is the closest to the true dose metric. This approach is unfortunately somewhat circular in reasoning, and exceptions to the use of the highest effect estimate have recently been shown.⁵⁵ Nevertheless, given all the uncertainties involved in observational retrospective epidemiology, it is ultimately in the exposure-response analysis that the validity of the exposure assessment is shown.

Discussion

Methods of assessment of exposure have been given much more attention in recent years. As a result, increasingly sophisticated approaches to retrospective assessment have been developed. There is also a growing recognition of the power inherent in building statistical models to describe factors associated with exposure. Despite improved methods, retrospective assessment of exposure remains hampered by the lack of complete and valid exposure data, especially for important historical periods.

Although many researchers have described

the estimation process in a multidimensional format, the generic concept of an EDM as a way of organising information related to exposure and providing the structure for estimation, including interpolation and extrapolation, has not been generally described. By considering exposure information in this structure, a wide variety of types of exposure information may be explicitly related, and various means of analysis considered. Almost any problem of exposure estimation can be structured in an EDM. For instance, the job exposure matrix (JEM) used to assign exposure in population based case-control studies is a closely related concept,⁵⁶ in which the available data are generally limited to the subjective assignment of potential exposures based on a job classification scheme across many industries. The EDM extends this basic approach to a wide variety of study contexts, data types, and analytical approaches.

Two areas remain especially challenging in methods of exposure assessment. To predict health effects validly, the methods adopted must be specifically tailored to the outcome under study. This need pertains to all aspects of the assessment process beginning with identifying the pertinent form and temporal characteristics of the agent measured, to the estimation of the appropriate variable of the exposure distribution, to the development of an appropriate and specific exposure metric and to the form of analysis to assess the outcome. Although some researchers have suggested that specific background knowledge of the disease mechanism is required for successful assessment of epidemiological exposure,^{21 55} others have relied on identifying the best fit model to the data to determine the form of the exposure assessment,²⁰ and still others have begun to adopt an integrated approach to exposure assessment and exposure-response analyses.⁵⁷

Secondly, a thorough understanding of the errors inherent in alternative approaches to exposure assessment, and their effects on exposure-response relations will provide a most important area for additional research and progress. The effect of measurement error or exposure misclassification has become widely acknowledged in the epidemiological literature. The specific ways in which alternative exposure assessment methods either contribute to, or help control these errors, are only beginning to be understood. The analysis and control of exposure errors are particularly important in studies of early or subtle health effects, or low levels of exposures. With the accumulation of a greater quantity and specificity of exposure data in recent decades, and the increasing use of prospective studies, the opportunity to exploit methods that explicitly consider effects of error in measurement will also emerge.

These two areas, integration of mechanistic understanding with statistical modelling and the understanding, control, and adjustment of errors in exposure assessment, will provide the greatest opportunities for progress in exposure assessment for occupational epidemiology. Progress on these issues will prove particularly important as epidemiological science is called on to quantitatively consider emerging issues

such as biomechanical stressors and electromagnetic fields. Nevertheless, no amount of foresight and prospective monitoring will replace the need for sound approaches to retrospective estimation of exposure, and the variety of methods now available provide a basis for that work.

- Stewart PA, Herrick R. International workshop on retrospective exposure assessment for occupational epidemiologic studies. *Appl Occup Environ Hyg* 1991;6:403-560.
- Rappaport SM, Smith TJ. *Exposure assessment for epidemiology and hazard control, industrial hygiene science series*. Chelsea, MI: Lewis 1991:313.
- Checkoway H, Pearce NE, Crawford-Brown DJ. *Research methods in occupational epidemiology*. New York: Oxford University Press, 1989:20.
- Checkoway H, Seixas NS, Demers PA. The influence of occupational exposure assessment on epidemiologic inferences. *Occup Hyg* 1995 (in press).
- Boleij J, Heederik D, Kromhout H. *Occupational exposure assessment*. Rotterdam: Elsevier, 1995.
- Seixas NS. Exposure assessment methods: environmental monitoring. In: Harber P, Schenker M, Balmes J, eds. *Occupational and environmental respiratory disease*. St Louis: Mosby Year Book, 1995:248-87.
- Ott MK, Norwood SK, Cook RR. The collection and management of occupational exposure data. *American Statistician* 1985;39:432-6.
- Klaassen CK, Doull J. Evaluation of safety: toxicologic evaluation. In: Doull J, Klaassen CD, Amdur MO, eds. *Casarett and Doull's toxicology, 2nd ed*. New York: Macmillan 1980:11-27.
- Thomas D, Stram D, Dwyer J. Exposure measurement error: influence on exposure-disease relationships and methods of correction. *Annual Reviews of Public Health* 1993;14:69-93.
- Armstrong BK, White E, Saracci R. *Principles of exposure measurement in epidemiology*. Oxford: Oxford University Press, 1992.
- Byar DP, Gail MH. Errors-in-variables workshop. *Statistics in medicine, vol 8*. New York: John Wiley, 1989.
- Carroll RJ, Stefanski LA, Ruppert D. *Nonlinear measurement error models*. London: Chapman and Hall, 1995.
- Ulvason U. Limitations to the use of employee exposure data on air contaminants in epidemiologic studies. *Int Arch Occup Environ Health* 1983;52:285-300.
- Gamble J, Spirtas B. Job classification and utilisation of complete work histories in occupational epidemiology. *J Occup Med* 1976;18:399-404.
- 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.
- Buringh E, Lanting R. Exposure variability in the workplace: its implications for the assessment of compliance. *Am Ind Hyg Assoc J* 1991;52:6-13.
- Rappaport SM, Kromhout H, Symanski E. Variation of exposure between workers in homogeneous exposure groups. *Am Ind Hyg Assoc J* 1993;54:654-62.
- Stewart WF, Stewart PA. Occupational case-control studies: I. Collecting information on work histories and work-related exposures. *Am J Ind Med* 1994;26:297-312.
- Stewart PA, Blair A, Dosemeci M, Gomez M. Collection of exposure data for retrospective occupational epidemiologic studies. *Appl Occup Environ Hyg* 1991;6:280-9.
- Seixas NS, Robins TG, Becker M. A novel approach to the characterisation of cumulative exposure for the study of chronic occupational disease. *Am J Epidemiol* 1993;137:463-71.
- Smith TJ. Exposure assessment for occupational epidemiology. *Am J Ind Med* 1987;12:249-68.
- Hawkins NC, Norwood SK, Rock JC. *A strategy for occupational exposure assessment*. Akron, OH: Am Ind Hyg Assoc, 1991.
- Corn M, Esmen NA. Workplace exposure zones for classification of employee exposure to physical and chemical agents. *Am Ind Hyg Assoc J* 1979;40:47-57.
- Seixas NS, Sheppard L. Maximising accuracy and precision using individual and grouped exposure assessment. *Scand J Work Environ Health* 1995; (in press).
- Nicas M, Spear RC. A task-based statistical model of a worker's exposure distribution: part I—description of the model. *Am Ind Hyg Assoc J* 1993;54:211-20.
- Olsen E. Analysis of exposure using a logbook method. *Appl Occup Environ Hyg* 1994;9:712-22.
- Esmen N. Retrospective industrial hygiene surveys. *Am Ind Hyg Assoc J* 1979;40:58-65.
- Seixas NS, Robins TG, Rice CH, Moulton LH. Assessment of potential biases in the application of MSHA respirable coal mine dust data to an epidemiologic study. *Am Ind Hyg Assoc J* 1990;51:534-40.
- Seixas NS, Moulton LH, Robins RG, et al. Estimation of cumulative exposures for the national study of coal workers pneumoconiosis. *Appl Occup Environ Hyg* 1991;6:1032-41.
- Kromhout H, Oostendorp Y, Heederik D, Boleij J. Agreement between qualitative exposure estimates and quantitative exposure measurements. *Am J Ind Med* 1987;12:551-62.
- Teschke K, Hertzman C, Dimich-Ward H, et al. A comparison of exposure estimates by worker raters and industrial hygienists. *Scand J Work Environ Health* 1989;15:424-9.
- Hawkins NC, Evans JS. Subjective estimation of toluene exposure: a calibration study of industrial hygienists. *Appl Ind Hyg* 1989;4:61-8.
- Post W, Kromhout H, Heederik D, Noy R. Semiquantitative estimates of exposure to methylene chloride and styrene: the influence of quantitative exposure data. *Appl Occup Environ Hyg* 1991;3:197-204.
- Rinsky RA, Smith AB, Hornung R, et al. Benzene and leukemia: an epidemiologic risk assessment. *N Engl J Med* 1987;316:1044-50.
- Attfield MD, Moring K. The derivation of estimated dust exposures for US coal miners working before 1970. *Am Ind Hyg Assoc J* 1992;53:248-55.
- Stewart PA, Blair A, Cubit D, et al. Estimating historical exposures to formaldehyde in a retrospective mortality study. *Appl Ind Hyg* 1986;1:34-41.
- Checkoway H, Heyer NJ, Demers PA, Breslow NE. Mortality among workers in the diatomaceous earth industry. *Br J Ind Med* 1993;50:586-97.
- Dodgson J, Cherie J, Groat S. Estimates of past exposure to respirable man-made mineral fibres in the European insulation wool industry. *Ann Occup Hyg* 1987;31:567-82.
- Cherie J, Krantz S, Schneider T, et al. An experimental simulation of an early rock wool/slag wool production process. *Ann Occup Hyg* 1987;31:583-93.
- Ayer HE, Dement JM, Busch KA. A monumental study: reconstruction of a 1920 granite shed. *Am Ind Hyg Assoc J* 1973;34:206-11.
- Dement JM, Harris RL, Symons MJ, Shy CM. Exposures and mortality among chrysotile asbestos workers. Part I: exposure estimates. *Am J Ind Med* 1983;4:399-419.
- Hallock MF, Smith TJ, Woskie S, Hammond S. Estimation of historical exposures to machining fluids in the automotive industry. *Am J Ind Med* 1994;26:621-34.
- Hammond SK. The use of markers to measure exposures to complex mixtures. In: Rappaport SM, Smith TJ, eds. *Exposure assessment for epidemiology and hazard control*. Chelsea: Lewis, 1991:53-66.
- Armstrong BG, Tremblay C, Cyr D, Theriault G. Estimating the relationship between exposure to tar volatiles and the incidence of bladder cancer in aluminium smelter workers. *Scand J Work Environ Health* 1986;12:486-93.
- Robins TG, Seixas NS, Franzblau A, Burge H. Respiratory effects of machining fluid aerosols. Final report: UAW-GM Occupational Health Advisory Board. Detroit, MI: OHAB, 1994.
- Greife AL, Hornung R, Stayner L, Steenland K. Development of a model for use in estimating exposure to ethylene oxide in a retrospective cohort mortality study. *Scand J Work Environ Health* 1988;14:29-30.
- Eisen EA, Smith TJ, Wegman DH, et al. Estimation of long term dust exposures in the Vermont granite sheds. *Am Ind Hyg Assoc J* 1984;45:89-94.
- Hornung RW, Greife AL, Stayner LT, et al. Statistical model for prediction of retrospective exposure to ethylene oxide in an occupational mortality study. *Am J Ind Med* 1994;25:825-36.
- Woskie SR, Smith TJ, Hammond S, Hallock M. Factors affecting worker exposures to metal-working fluids during automotive component manufacturing. *Appl Occup Ind Hyg* 1994;9:612-21.
- Teschke K, Marion S, van Zuylen M, Kennedy S. Maintenance of stellite and tungsten carbide saw tips: determinants of exposure to cobalt and chromium. *Am Ind Hyg Assoc J* 1995;56:661-9.
- Hornung RW. Statistical evaluation of exposure assessment strategies. *Appl Occup Environ Hyg* 1989;6:516-20.
- Dosemeci M, McLaughlin J, Chen J, et al. Indirect validation of retrospective exposure assessment method used in a nested case-control study of lung cancer and silica exposure. *Occup Environ Med* 1994;51:136-8.
- Kriebel D. The dosimetric model in occupational and environmental epidemiology. *Occup Hyg* 1994;1:55-68.
- Salvan A, Stayner L, Steenland K, Smith R. Selecting and exposure lag period. *Epidemiology* 1995;6:387-90.
- Kauppinen T, Partanen T. Use of plant- and period-specific job-exposure matrices in studies on occupational cancer. *Scand J Work Environ Health* 1988;14:161-7.
- Ballew MA, Kriebel D, Smith TJ. Epidemiologic application of a dosimetric model of dust overload. *Am J Epidemiol* 1995;141:690-6.