Applied Occupational and Environmental Hygiene

ISSN: 1047-322X (Print) 1521-0898 (Online) Journal homepage: https://www.tandfonline.com/loi/uaoh20

Statistical Evaluation of Exposure Assessment Strategies

Richard W. Hornung

To cite this article: Richard W. Hornung (1991) Statistical Evaluation of Exposure Assessment Strategies, Applied Occupational and Environmental Hygiene, 6:6, 516-520, DOI: 10.1080/1047322X.1991.10387921

To link to this article: https://doi.org/10.1080/1047322X.1991.10387921

Statistical Evaluation of Exposure Assessment Strategies

Richard W. Hornung
National Institute for Occupational Safety and Health, 4676 Columbia Parkway, Cincinnati, Ohio 45226

It is well known that exposure measurement error or misclassification can bias an epidemiologic risk assessment. Accordingly, it is useful to have methods of assessing the degree of error in exposure assessments. This paper proposes statistical evaluation techniques for two types of exposure assessment strategies: categorization of exposure and estimation of actual personal exposure levels (with particular attention to the latter).

Categorization of exposure is generally done in one of the two ways: ordinal but nonquantitative grouping or quantitative exposure intervals. Estimation of actual exposure levels is most often done in the form of job-by-year exposure matrices and somewhat less often by exposure prediction models. The nature of uncertainty inherent with each of these approaches is described.

Quantitative estimates of the magnitude of error in an exposure assessment are possible when exposure matrices or prediction models are employed. Three types of statistical approaches are explored for estimation of the degree of error in quantitative exposure assessments, with particular attention to methods associated with exposure prediction models. The methods described are sensitivity analysis, cross-validation techniques, and field testing. Advantages and disadvantages of each method are discussed, as well as the appropriate situation for their application. Cross-validation and sensitivity analysis are shown to be particularly well-suited to the validation of exposure prediction models. Field testing is a more general technique that can be used to assess the validity of exposure matrices. **Hornung, R.W.: Statistical Evaluation of Exposure Assessment Strategies. Appl. Occup. Environ. Hyg. 6:516–520; 1991.**

Introduction

Two of the most critical aspects regarding the credibility of an epidemiologic risk assessment are the effect on risk estimates of errors in exposure indices and the evaluation of the nature and extent of such errors in the exposure assessment. There has been considerable attention in the literature to the former problem, but very little has been done regarding statistical methods for evaluation of exposure assessments. The purpose of this paper is to suggest several methods based upon existing statistical techniques to estimate the bias and precision associated with exposure estimates.

The statistical methods used for evaluation depend to some degree upon the particular exposure assessment strategies used. There are several different strategies commonly reported for assigning exposure estimates to members of an epidemiologic study group. These strategies can generally be divided into two classes: categorization of exposure and estimation of actual personal exposure levels. Exposure categorization includes the simple dichotomy of exposed versus unexposed and the ordinal categories corresponding to such designations as unexposed, low, moderate, and high exposures. Estimates of actual exposure levels are usually in the form of a job-by-year exposure matrix or exposure prediction models. The evaluation methods described in this paper will be directed principally toward exposure prediction models; however, some of them can be applied to any exposure assessment strategy.

Sources of Uncertainty

Before attempting to develop evaluation criteria for various exposure assessment strategies, it is important to examine the sources of uncertainty inherent in each approach. Since a simple dichotomy involving exposed versus unexposed persons is of little use to risk assessors, three strategies will be considered: ordinal classification, exposure matrices, and exposure prediction models.

Ordinal Classification

Ordinal classification is a technique widely used, especially when the epidemiologic study is analyzed using life table analyses. The classification can be done in two ways. The first involves the determination of the degree of exposure for each study member based upon limited exposure data, knowledge of job characteristics, and proximity to the exposure source. This technique is typically used when little, if any, actual exposure measurement data are available for most study members. When this method is used, it is usually not possible to assign a numerical value for exposure to each category. An assumption is made that each successive category involves higher ex-

posures than the previous one, e.g., low, medium, and high, without any exposure level designated for each. The second method of ordinal classification divides the total range of exposure into mutually exclusive intervals based upon measured or hypothesized quantitative exposure levels. This method has the advantage that quantitative values can be assigned (usually the midpoint) to each person in a given interval.

The sources of uncertainty associated with ordinal classification are potentially serious. If the classification is done using exposure potential, there is no real basis for estimation of a quantitative risk assessment. Even if each category is given a score, one can only conclude that a trend may exist in disease risk with increasing exposure potential. This may help to establish causality but will not estimate a level of risk, given that a person is exposed to a measured value of some contaminant. Because no quantitative exposure levels are assigned to each exposure classification, the only method for evaluating such an exposure assessment is to verify that each exposure scenario is indeed in the right order. That is, if a given person's work history is classified or scored as "moderate exposure," that person, on average, would have experienced exposures greater than those persons in the "low" category but less than those persons in the "high" category. Since this method of exposure assessment is generally used when few, if any, exposure measurement data are available, the degree of uncertainty is high. Statistical techniques cannot be applied to this exposure assessment method unless at least some monitoring data are available. Assessment of the appropriateness of classification is probably best done by a panel of expert industrial hygienists familiar with the exposures being studied. The degree of uncertainty associated with ordinal classifications based upon exposure level intervals is generally less than categorization of exposure potential. At the very least, the numerical values for exposure in each interval make possible a statistical evaluation of uncertainty.

One of the most important sources of uncertainty when using ordinal classification based upon exposure level intervals involves the selection of cutpoints. It has been shown by Morgan and Elashoff,[1] among others, that categorization of a continuous covariate can increase the variance of risk estimates derived from such exposure data. In addition, improper choice of such intervals can increase misclassification of exposure, especially when natural boundaries for exposure levels are ignored. For example, if one is attempting to assess exposure to cigarette smoking, careful attention must be given to the choice of exposure categories for cigarettes smoked per day. This is due to the fact that most persons report smoking in units of packs per day. In general, it is better to use continuous measures of exposure rather than an arbitrary categorization. This is particularly true when the alternative is to assign a broad range of job titles to each of a small number of exposure categories. An exception to this rule is encountered when the method of analysis for the associated epidemiologic study involves the necessity of grouped data, e.g., life table analysis.

Exposure Matrices

Exposure matrices are usually created using a cross-classification of exposure factors such as job, department, plant, and calendar year. If extensive data are available, each cell of such a matrix will consist of some summary measure of exposure (generally, the mean) for all exposure measurements made on that combination of factors. In general, this is the preferred method of exposure assessment when the measurement data are extensive enough to cover all such combinations. When this is the case, the only source of uncertainty is sampling and analytical error, which is minimal. This situation, however, is seldom encountered in practice.

A more common occurrence is the absence of measurement data for a substantial number of cells in the exposure matrix. This is particularly true for the earlier calendar years needed for a retrospective epidemiologic study. When creating an exposure matrix, the level of uncertainty rises with the number and distribution of empty data cells. The degree of uncertainty is also a function of the manner in which the exposure levels are estimated for the missing cells. Interpolation is an accepted method of dealing with missing cells, especially between adjacent time intervals. Extrapolation over time, on the other hand, will raise the level of uncertainty in the absence of knowledge concerning other important exposure factors. The method selected for estimating exposure levels for missing cells in an exposure matrix leads naturally into the discussion of exposure prediction models.

Exposure Prediction Models

An exposure prediction model is any mathematical expression that generates an estimate of exposure levels as a function of one or more input variables. These input variables are usually such exposure factors as job, department, location, calendar year, or use of engineering controls. Some exposure prediction models are deterministic expressions which may have been produced using principles of physics or laboratory studies. An example of such a model is the air dispersion model used by the U.S. Environmental Protection Agency (EPA) for estimating environmental exposures to populations living in the vicinity of an airborne emission source.[2]

Another type of exposure prediction model is created by actually fitting a statistical model to existing industrial hygiene data including exposure measurements and data regarding important exposure factors. An example of this type of model is illustrated in an exposure assessment of ethylene oxide in the sterilization industry.[3]

Sources of uncertainty in exposure prediction models involve three areas, the first two of which also apply to exposure matrices. First, there is the usual degree of sampling and analytical error inherent in the measurement data when a statistical model is used. This may be compounded if the data were not collected by means of a probability sample, e.g., random sampling. Since random sampling is seldom used for collection of exposure data,

the degree of uncertainty depends upon how "representative" the measurement data are relative to the actual population of all exposure conditions. This may be a serious problem if most measurements were taken only when conditions were expected to be atypical. A second source of uncertainty concerns the accuracy with which data are available concerning the exposure factors. Certain input variables such as calendar year and job type may be known without error, but others, such as production levels, ventilation, and the extent of use of other engineering controls, may be only unreliable estimates. The third source of uncertainty is the specification of the model itself. If the model is incorrectly formulated, predicted exposure levels can be expected to be biased either too high or too low. This type of uncertainty is often encountered when attempting to specify the form of the model during periods of time when no exposure data and/or information on exposure factors exist.

The next section addresses several statistical methods that can be used to evaluate the accuracy of exposure assessments. The main focus of this section deals with validation of exposure prediction models; however, many of these techniques can be used with ordinal classification or exposure matrices.

Statistical Evaluation Methods

The determination of which statistical methods are appropriate for estimation of uncertainty depends upon the parameters being estimated, the nature and amount of data available, and the estimation method being used. The primary methods discussed in this paper include sensitivity analysis, cross-validation techniques, and field testing. In general, one assumes that the expected value, or mean exposure level, is the parameter of interest. Of secondary interest is the variation, or equivalently, the standard error of the estimated mean exposure level.

Sensitivity Analysis

Sensitivity analysis is a statistical technique by which exposure factors are identified as having the maximum effect upon the exposure estimates. This technique is valuable in quantification of the expected change in predicted exposure levels as a function of the range of values for each input variable. In this way, individual exposure factors can be identified that are capable of driving the exposure prediction model. There are two principle methods for conducting a sensitivity analysis. The first is the easier of the two but is less informative.

The first technique involves standardization of the coefficients for all input variables by dividing each coefficient by the ratio of the standard deviation of the measured exposure level to the standard deviation of each input variable. Since the denominators cancel out, this can be written as:

$$\beta_j^* = \frac{\beta_j}{\sqrt{[\sum (Y_i - \bar{Y})^2]/[\sum (X_{ij} - \bar{X}_j)^2]}} \quad (1)$$

$$= \beta_j \sqrt{\sum (X_{ij} - \bar{X}_j)^2 / \sum (Y_i - \bar{Y})^2}$$

where: β_j^* = standardized coefficient for jth exposure factor
β_j = usual regression coefficient for jth exposure factor
Y_i = measured exposure for ith worker
X_{ij} = value of jth exposure factor for ith worker

If multiple regression techniques are used to develop the exposure prediction models, many statistical packages, e.g., SAS, have options to calculate these standardized coefficients. The square of the standardized coefficient (β_j^{*2}) can be interpreted as the fraction of the total variation in exposure levels attributable to the jth input variable. In this way, the relative importance of each of the exposure factors can be determined. However, because of correlation among these factors, this is only an approximate estimate of the relative strength of each factor individually. A more useful, although more time-consuming, technique involves the calculation of the degree of change in exposure estimates as a function of the range of each input variable. The usual method for this type of sensitivity analysis is to determine the probable minimum and maximum values for each exposure factor and then calculate the estimated exposure level for each while holding all other exposure factors at their mean value. This method not only identifies a relative ordering of the importance of each exposure factor, but it also identifies potential errors in the specification of the exposure prediction model. If the predicted exposure at either the maximum or minimum value for a given exposure factor is unrealistic, the model should be modified in some way. Another, potentially more difficult, sensitivity analysis involves an investigation of the global maximum and minimum values for the estimated exposure levels. This is done by creating a best- and worst-case scenario involving all exposure factors simultaneously. That is, each exposure factor is set at the level that produces either minimum or maximum estimated exposures. Care must be taken that the selected minimum or maximum values for each exposure factor have a positive probability of occurring when the overall exposure assessment is done. For example, one would avoid setting the production level at its highest value in combination with a calendar year where that level of production was never reached.

This latter type of sensitivity analysis identifies the range of possible predicted exposures. When this range is unrealistic based upon measured data or a simple understanding of the process, it may be necessary to restrict or modify the model in some way to produce a reasonable range of predicted values. This may be particularly true for extrapolations based on values for the input variables that were outside the range of such variables used in developing the model. An exposure factor particularly sensitive to extrapolation errors is calendar time. If there is a trend in exposures over time even after adjusting for changes in such factors as engineering controls and production levels, caution should be taken in prediction of exposures occurring many years before any exposure measurements were available.

Cross-Validation Techniques

Cross validation is the statistical term for a validation method by which models are developed using one subset of a data set while the other subset is used for evaluating the accuracy of the model. Generally, the subset of the data used for validating the model is considerably less than half of the full data set. In this way, little precision is lost in estimating the coefficients for the model compared to using the full data set for model development. When the full data set is used to develop the model, the natural estimate of the variance of the predicted values would seem to be the mean square error for the model:

$$MSE = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n - p} \quad (2)$$

where: $\hat{y}_i$ = the predicted exposure
y_i = the observed exposure measurement
n = the number of observations
p = the number of parameters estimated in the model

However, this has been shown to underestimate the true error rate for prediction of exposures not used to develop the model.[4,5] This biased estimate of the true error is due to the fact that the model was developed (often by least squares) to fit the existing data as closely as possible. The magnitude of the bias is a decreasing function of the number of data points available to develop the model and an increasing function of the number of parameters estimated in the model. Because of this underestimate of the model's true error rate, comparison of the model to the validation subset of data provides a better measure of the model's accuracy.

Most of the statistical theory regarding cross-validation techniques assumes a probability sample of the underlying population of measurement values. The theory further assumes that subsequent predictions using the fitted model are estimates of mean predicted values for the same population distribution. In practice, one is often not sure that future predictions involve the same conditions under which the sample measurements were taken. Under theoretical conditions, the portion of the data withheld from model development in order to validate the model is usually a random sample of the original data set. Several papers have been written advocating a nonrandom sample such that the total range of the input parameters used in developing the model is approximated by the range of all input variables in the validation set.[5,6] It stands to reason that selection of a validation subset of the measurement data where there is some doubt about its representativeness compared to the input variables used in the model development will provide the most severe test of the model. As an example, in the validation of the ethylene oxide prediction model, data from several plants were withheld from model development. Conditions in these plants may have been somewhat different than in those plants for which data was used to develop the model. Therefore, the use of these data probably resulted in a greater (and more realistic) estimate of model error than an estimate generated by use of a random sample across all plants in the study.

Accuracy is characterized by two components: precision and bias. Precision is a measure of the degree of variability of the estimate of a true exposure level. Bias is the average difference between the estimates and the true exposure levels. In order to estimate bias directly from the validation data set, it is necessary to assume that these values are essentially measured without error. There is an obvious problem with this approach since the measurements used for validation are also subject to inaccuracy and environmental variability. Nevertheless, use of an independent source of data provides some idea of the variability and bias associated with the prediction model.

The estimates of precision and bias are calculated by taking the standard deviation and the mean of the individual differences between the validation measurements and the predicted exposure levels. This can be expressed as:

$$bias = \sum_{i=1}^{n_o} \frac{(y_i - \hat{y}_i)}{n_o} \quad (3)$$

$$precision = \sqrt{\sum_{i=1}^{n_o} \frac{[(y_i - \hat{y}_i) - bias]^2}{n_o - 1}} \quad (4)$$

where: $\hat{y}_i$ = predicted exposure level for the ith set of exposure factors in the validation data set
y_i = the measured exposure for the ith set of exposure factors
n_o = number of observations in the validation data set

The overall estimate of accuracy is then obtained by combining the estimated bias and precision in a mean-squared sense:

$$Accuracy = \sqrt{(precision)^2 + (bias)^2} \quad (5)$$

The smaller the value of this statistic, the more accurate is the estimated exposure. Although this is the usual method for calculating total error in a components of variance application, the EPA has also suggested an alternative measure. They have used as a measure of total error the relative bias plus twice the coefficient of variation. This can be expressed as:

$$Total\ error = \frac{bias}{\bar{y}} + 2\frac{precision}{\bar{y}} \quad (6)$$

The EPA then specifies that the total error in the prediction model is acceptable if Equation 6 is less than 50 percent of the mean exposure level.[7] This requirement probably is unnecessarily strict for prediction of exposures in occupational studies where day-to-day variability alone may often exceed 50 percent of the mean concentration level. In such studies, the acceptable total error should probably be set at some multiple of the combined sampling, analytical, and short-term temporal variability of the measured exposures.

Field Testing

When limited data are available for use with the cross-validation technique or when prediction models are not

developed from measurement data, an alternative method for assessing uncertainty is field testing. This is the most direct measure of validity of an exposure assessment, at least under current exposure conditions. It is a straightforward technique by which the prediction of exposure levels is compared to actual measurements made during an industrial hygiene survey specifically designed to test the exposure estimation method. This type of validation has the most general application since it can be used to test any form of exposure assessment including exposure matrices. The major disadvantage of this method is that many historical exposure conditions cannot be replicated under current operational situations.

Validation of an exposure assessment using field testing is conceptually very simple. The exposure factors determined to affect exposure levels in either a model or a job exposure matrix are listed together with their associated ranges for use in the retrospective health study. Exposure sites are located which cover as many of these ranges as possible. In certain situations, the value of the exposure factor may be controlled or manipulated to mimic the set of conditions in the retrospective health study. This form of experimental design was recently applied to a study of formaldehyde levels in the embalming industry by National Institute for Occupational Safety and Health (NIOSH) and National Cancer Institute (NCI) investigators. They were able to control such exposure factors as ventilation rates, solution strengths, and condition of the body being embalmed. The model generated from these experimental data is in the process of validation using field testing in funeral homes.

Summary

The type of technique used for validation of an exposure assessment depends to some extent upon the type of assessment methodology being used. This paper deals principally with exposure prediction models, although some of the methods discussed may be applicable to virtually any type of exposure assessment.

Sensitivity analysis can be used in two ways: 1) to determine the exposure factors having the strongest effect upon the exposure level and 2) to identify combinations of exposure factors that predict exposure levels which are out of any reasonable range. This technique is meant to be used for exposure prediction models either developed from measurement data or by some deterministic method.

Cross-validation methods would seem to be the most valuable technique for evaluating prediction models developed from exposure measurements. The most severe test of such models, i.e., the one least likely to favor the model, involves selection of a subset of the data encompassing different plants or calendar years for which the homogeneity of the other exposure factors is not assured. Random samples for validation purposes are statistically valid but are more likely to agree with the model prediction, especially when the original measurement data did not arise from a probability sample.

Field testing is more generally applicable to validation of any type of exposure assessment. By designing a survey to take samples under exposure conditions used in either a prediction model or an exposure matrix, the uncertainty of the exposure assessment can be quantified. The disadvantages of this approach are that it is expensive and it may be limited only to current exposure conditions. It is most valuable when current conditions can be altered to reflect historical exposure factors that occurred during the period of the retrospective health study.

References

1. Morgan, T.M.; Elashoff, R.M.: Effect of Categorizing a Continuous Covariate on the Comparison of Survival Time. J. Am. Stat. Assoc. 81(396):917–921 (1986).
2. Turner, D.B.: Workbook of Atmospheric Dispersion Estimates. U.S. Department of Health, Education, and Welfare, Cincinnati, OH (1969).
3. Greife, A.L.; Hornung, R.W.; Stayner, L.G.; Steenland, K.N.: Development of a Model for Use in Estimating Exposure to Ethylene Oxide in a Retrospective Cohort Mortality Study. Scand. J. Work Environ. Health 14(Suppl):29–30 (1988).
4. Efron, B.: How Biased is the Apparent Error Rate of a Prediction Rule? J. Am. Stat. Assoc. 81(394):461–470 (1986).
5. Picard, R.R.; Cook, R.D.: Cross-Validation of Regression Models. J. Am. Stat. Assoc. 79(387):575–583 (1984).
6. Snee, R.D.: Validation of Regression Models: Methods and Examples. Technometrics 19:415–428 (1977).
7. Pierson, S.; Lucas, R.: Hispanic HANES Pilot Study: Measurement of Volatile and Semivolatile Organic Compounds in Blood and Urine Specimens. EPA 560/5-83-001. U.S. Environmental Protection Agency, Washington, DC (1983).