Significance of Genetic Information in Risk Assessment and Individual Classification Using Silicosis as a Case Model

ERIN McCANLIES1*, DOUGLAS P. LANDSITTEL1, BERRAN YUCESOY2,4, VAL VALLYATHAN3, MICHAEL L. LUSTER2 and DAN S. SHARP1

1Biostatistics Branch, 2Toxicology and Molecular Biology Branch and 3Pathology and Physiology Research Branch, Health Effects Laboratory Division, National Institute for Occupational Safety and Health, 1095 Willowdale Road, Morgantown, WV 26505, USA; 4Ankara University, School of Pharmacy, Department of Toxicology, 06100 Ankara, Turkey

Received 16 October 2001; in final form 7 February 2002

Over the last decade the role of genetic data in epidemiological research has expanded considerably. We recently published a case–control study that evaluated the interaction between silica exposure and minor variants in the genes coding for interleukin-1α (IL-1α), interleukin-1 receptor antagonist (IL-1RA) and tumor necrosis factor α (TNFα) as risk factors associated with silicosis, a fibrotic lung disease. In contrast, this report uses data generated from these studies to illustrate the utility of genetic information for the purposes of risk assessment and clinical prediction. Specifically, this study will address how, given a known exposure, genetic information affects the characterization of risk groups. Relative operating characteristic (ROC) curves were then used to determine the impact of genetic information on individual classification. Logistic regression modeling procedures were used to estimate the predicted probability of developing silicosis. This probability was then used to construct predicted risk deciles, first for a model with occupational exposure only and then for a model containing occupational exposure and genetic main effects and interactions. Results indicate that the exposure-only model effectively captures an increasing relationship between predicted risk deciles and prevalence of observed silicosis cases. Individuals comprising the highest risk decile were almost four times as likely to have silicosis as opposed to the lowest risk decile. The addition of genetic data, however, substantially improved characterization of risk categories; the proportion of cases in the highest risk decile was almost eight times that in the lowest risk decile. However, the ROC curve and classification analysis demonstrated that the addition of genetic main effects and interactions did not significantly impact on prediction of the individual’s case status. These results indicate that genetic information plays a valuable role in effectively characterizing risk groups and mechanisms of disease operating in a substantial proportion of the population. However, in the case of fibrotic lung disease caused by silica exposure, information about the presence or absence of the minor variants of IL-1α, IL-1RA and TNFα is unlikely to be a useful tool for individual classification.

Keywords: epidemiology; genetics; occupational safety and health; silicosis; risk assessment

INTRODUCTION

Research about the factors associated with occupational diseases and hazards considers how characteristics of the host interact with aspects of the environment to influence risk. This typically consists of collecting data associated with exposure, disease and host characteristics such as age, race and gender. While observational studies are most often utilized to identify risk factors associated with disease, the exploration of genetic information holds the possibility of better understanding mechanisms of causation, delineating gene–environment interactions and evaluating individual susceptibility. Indeed, pharmacogenomic data provides opportunities to account for genetic
variability in order to improve drug treatment, while measures of DNA adduct formation are being used to better understand biological effects of specific exposures, to identify specific risk factors and to elucidate mechanisms of causation (Greim et al., 1995; Harris, 1995; Portier and Bell, 1998; Kuhlmann, 1999; Stephens, 1999). The use of genetic information to improve public health measures for the treatment and prevention of disease is an active area of scientific and public debate (Perera, 1996, 2000). This is true not only for infectious and chronic diseases, but also in the field of occupational safety and health.

Silica is an abundant and widespread mineral, primarily found in nature as silicon dioxide. Exposure to silicon compounds at levels high enough to result in silicosis, a chronic fibrotic lung disease, occurs in numerous occupations such as mining, tunneling, sandblasting, the ceramics industry and the manufacture of paint (Fraser et al., 1994; Wagner, 1997; National Institute for Occupational Safety and Health, 1992). In the USA between 1987 and 1999 between 1629 and 2599 individuals per year died of silicosis or coal workers’ pneumoconiosis (CWP) (National Institute for Occupational Safety and Health, 1999). Between 1988 and 1992 it was estimated that >1000000 US workers were exposed to silica, with ~59000 of these workers at risk of developing silicosis (National Institute for Occupational Safety and Health, 1992, 1994). These statistics strongly indicate the public health impact this disease has on workers and the need to better identify methods of treatment and prevention. To accomplish this, risk factors that, after intervention, would most effectively reduce disease must be identified.

In the current study we have utilized data from two recently published case–control studies (Yucesoy et al., 2001a,b) to evaluate how, given silica exposure, genetic information affects the characterization of risk groups and of individuals. These previous studies were designed to investigate the role of minor variants coding for interleukin-1α (IL-1α), interleukin-1 receptor antagonist (IL-1RA) and tumor necrosis factor α (TNFα) in the risk of silicosis in underground miners (Yucesoy et al., 2001a,b). Genetic variation in the DNA coding for these cytokines has been implicated in a number of health conditions characterized by chronic inflammatory processes antecedent to clinical disease (Yucesoy et al., 2001a,b). Thus, the study may be reasonably representative of occupational diseases in which exposure is required but in which genetic variation has also been found to influence risk.

The data being utilized was limited in that the overall proportion of cases (~0.62) had been fixed by previous study design. Therefore, the proportion of cases and odds ratios are not interpretable as population risk, or even relative risk.

MATERIALS AND METHODS

Study population

A detailed description of the study population has been given previously (Yucesoy et al., 2001a,b). Briefly, a case–control study that utilized autopsy lung tissue from underground coal miners was conducted to evaluate the role of minor variants in the genes that code for IL-1α, IL-1RA and TNFα in the risk of fibrotic lung disease caused by silica. Three hundred and eighteen cases and 163 controls were selected from a subset of over 6580 lung tissue samples collected between 1972 and 1996 for the National Coal Workers Autopsy Study (NCWAS) (Yucesoy et al., 2001a,b). The cases with moderate and severe silicosis were selected and considered eligible for study participation. The control samples were limited to Caucasian men without pathological evidence of disease who had worked in underground mines. Due to a sampling error, 35 (12%) of the controls who died at <55 yr of age with <16 yr exposure were excluded from the study, leaving 163 eligible controls. However, as described in the previous manuscripts, this omission had no practical effect on the results (Yucesoy et al., 2001a,b). For the purposes of the current analyses data were limited to 238 silicosis cases and 149 controls, all of whom had complete genotype information.

Pathology

Samples were reviewed and graded for CWP, including silicosis and other diseases, according to a joint committee of the National Institute for Occupational Safety and Health and the College of American Pathologists (Kleinerman et al., 1979). CWP lesions, such as nodules, progressive massive fibrosis (PMF), macules and silicosis were subjectively classified into three grades of no disease, mild/moderate and severe. Details of the grading methods were previously described (Yucesoy et al., 2001a,b). All samples that met the criteria for a silicosis case and could be successfully genotyped were included in the study. For the purposes of the current paper genetic polymorphisms were limited to genes that demonstrated consistent relations with different severities of disease, and disease was treated as a simple dichotomous outcome, combining moderate and severe cases.

Genotyping

As previously described, genotyping was performed to evaluate the association between silicosis and the presence of single nucleotide polymorphisms (SNP) in IL-1α, IL-1RA and TNFα. Specifically, SNPs were evaluated in the promoter region at position −308 in the gene encoding for TNFα, a T→C transition at position +2018 in the second exon of IL-1RA and a C→T SNP at +4845 in the gene encoding for IL-1α (Yucesoy et al., 2001a,b).
Exposure

Occupational exposure history, smoking history and demographic information was collected through a questionnaire completed by the next of kin. Since the study was limited to underground miners, years of experience, as a continuous variable, was used as a surrogate measure for total cumulative exposure to silica.

Attempts were made to derive a more precise characterization of silica exposure using previous studies of dust exposure in miners (Attfield and Morling, 1992). However, estimates were severely limited by the sparsity of data prior to 1970, from which 77% of the current population had been derived. This sparsity reflected several factors, including (i) only a small number of mines having been sampled prior to 1970 (29 large mines and other smaller mines), and (ii) numerous job titles not being represented in any, or in very few, of these pre-1970 samples. For example, ~42% of the miners in the current study reported job titles represented by less than 100 total samples of dust exposure. Accordingly, total years of underground mining experience was the best available characterization of occupational silica exposure for the population under study.

Statistical methods

Logistic regression models were fitted to the data using a model with a linear term for occupational exposure as the only predictor and a model with occupational exposure as well as genetic main effects and interactions. The second model included main effect terms for occupational exposure, presence of IL-1α, IL-1RA and TNFα minor variants, the two-way interactions IL-1α × TNFα, IL-1RA × TNFα, IL-1α × exposure and TNFα × exposure and the three-way interaction IL-1α × TNFα × exposure. Using the same data set from the previously described study of genetic risk factors, this model was specified by examining likelihood ratio tests from appropriate study of genetic risk factors, this model was specified.

Using the same data set from the previously described study of genetic risk factors, this model was specified by examining likelihood ratio tests from appropriate study of genetic risk factors, this model was specified. Predicted probabilities of disease were then calculated for each individual using the logistic model with occupational exposure and genetic main effects and all previously defined interactions. The resulting predicted probabilities were then ranked and again categorized into predicted risk deciles. The exact interpretation of these predicted risk deciles is more complex, because the individual’s predicted risk is based on the joint distribution of occupational exposure, status of all three polymorphisms and interactions specified in the model. This strategy has been used in cardiovascular research to effectively assess the added joint contribution of groupings of variables (e.g. blood lipids) compared with other groupings of variables (e.g. blood markers of chronic inflammation) as risk factors for coronary events (Yarnell et al., 1991).

For each risk decile the relative proportion of cases was then calculated as the proportion of cases observed in that particular risk decile divided by the proportion of observed cases in the lowest risk decile. By definition, the relative proportion was equal to 1.0 for the lowest decile. For a given logistic model differences in the relative proportions of observed cases among successive predicted risk deciles provide an assessment of how effectively the model characterizes risk of silicosis. Instances where the relative proportions remain approximately constant would indicate that the model offers little or no utility in differentiating heterogeneity of risk. In comparison, if the relative proportions of observed cases increased substantially among successive predicted risk deciles, the logistic model and subsequent predicted risk deciles would then reflect a useful categorization of risk based on the covariates in the model.

Similarly, comparisons could be made between logistic models (with and without genetic data) by contrasting the degree to which the relative proportions of observed cases increase with increasing predicted risk deciles. The model with a steeper increase in relative proportions would thus be more useful in terms of risk assessment.

To assess classification accuracy on an individual level, the percent correctly classified and resulting ROC curves (Erdreich and Lee, 1981) were reported for each logistic model (with and without genetic data). ROC analysis is a method that facilitates the comparison of classification results between two different models. This is done by plotting the probability of true positives (sensitivity) versus the probability of false positives (1 – specificity) for various decision criteria. As classification accuracy depends on a specific cut-off point for predicting silicosis status, the ROC curve reflects classification accuracy over the entire range of prediction rules. The model
corresponding to the higher ROC curve would be considered more effective for classification.

The sensitivity and specificity characteristics reported in this study are determined by the study design with a fixed proportion of cases and thus should not be quantitatively generalized even to a population of coal miners. However, the relative features are sufficient to draw inferences about the role of genetic data in individual classification.

**RESULTS**

Complete data on disease, occupational exposure and gene status were obtained for 387 individuals, including 238 silicosis cases and 149 controls (Table 1). The cases were significantly older than the controls (68.5 versus 62.9 yr; \( P < 0.001 \)) and accordingly had greater numbers of years exposed to silica (33.6 versus 21.5 yr; \( P < 0.001 \)). Without adjusting for other covariates, minor variants of TNF\( \alpha \) (56 versus 52%; not significant), IL-1RA (49 versus 27%; \( P < 0.001 \)) and IL-1\( \alpha \) (22 versus 20%; not significant) all occurred more frequently in cases compared with controls (Table 1).

Occupational exposure alone led to identification of risk groups associated with increasing proportions of cases, especially for the lower half of the distribution (\( P < 0.01 \)) (Fig. 1). The proportions of cases in the third, fourth and fifth decile, respectively, were equal to 1.8, 2.8 and 3.1 times the proportion in the lowest decile. The trend was less pronounced in the upper half of the distribution as the relative proportions varied between 3.1 and 3.8, with subjects in the top decile having 3.5 times the proportion of silicosis cases as compared with the lowest decile.

The model with occupational exposure and genetic data showed a steeper relationship between predicted risk decile and the relative proportion of observed cases (Fig. 1). The resulting joint subset of genetic main effects and interactions with occupational exposure was statistically significant at \( P < 0.001 \), using the likelihood ratio test, when compared with the occupational exposure-only model. With one exception, the relative proportion progressively increased with each risk decile. Subjects in the highest risk decile were 7.6 times more likely to be a silicosis case. Even the

---

### Table 1. Descriptive characteristics of silicosis cases and controls

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>238</td>
<td>149</td>
</tr>
<tr>
<td>Age (± SD)</td>
<td>68.5 ± 9.0</td>
<td>62.9 ± 8.1*</td>
</tr>
<tr>
<td>Exposure (yr ± SD)</td>
<td>33.6 ± 11.1</td>
<td>21.5 ± 13.5*</td>
</tr>
<tr>
<td>TNF( \alpha ) (–308) (%) ( ^b )</td>
<td>134 (56.3)</td>
<td>78 (52.4)</td>
</tr>
<tr>
<td>IL-1( \alpha ) (+4845) (%) ( ^b )</td>
<td>52 (21.9)</td>
<td>30 (20.1)</td>
</tr>
<tr>
<td>IL-1RA (+2018) (%) ( ^b )</td>
<td>116 (48.7)</td>
<td>40 (26.9)*</td>
</tr>
</tbody>
</table>

*\( P \leq 0.001 \).

*Because the overall proportion of cases (~62%) was fixed by study design, reported frequencies do not represent population estimates. Percentages represent simple proportions of the minor variant; all three genetic polymorphisms were involved in statistically significant interactions with each other and/or occupational exposure and are therefore included in the analysis.
second decile was associated with a higher proportion of cases (2.6 times the proportion in the lowest decile). These results indicate that the joint addition of genetic factors into the model leads to more effective characterization for silicosis risk assessment.

Results of individual classification analyses, using the previously specified logistic models to predict individual probabilities of risk in conjunction with ROC analysis, indicate that addition of genetic covariates and interactions does not effectively increase accuracy (Fig. 2). Over a range of classification rules (i.e. using different cut-off points of predicted probability for classifying observations as a case) the percent correctly classified generally varied between 70 and 75% for both the exposure-only and exposure + genetics models. Sensitivity and specificity generally ranged between 70 and 90% and 50 and 70%, respectively, for either model. The curves appear indistinguishable over most of the range, with the addition of genetic data leading to slightly better prediction only over the lower, less meaningful part of the curve. Indeed, as a measure of relative effectiveness, the total area under the ROC curve was 0.745 for the exposure-only model and 0.775 for the model including genetic main effects and interactions. The region of these curves where the greatest separation was noted corresponds to classification results that are similar or worse than those that would be obtained by chance, further underscoring the similarity of the two models with respect to predicting individual case status.

**DISCUSSION**

Data from a previously published study of genetic variants enabled us to evaluate how genetic information in conjunction with occupational exposure data can be utilized to assess risk and predict silicosis case status. Addition of genetic data substantially improved characterization of risk categories, however, the ROC curve and classification analysis demonstrated that the addition of genetic information to exposure information did not effectively improve individual classification accuracy.

A principle limitation of these data with respect to estimation of population risk is that the overall proportion of cases was fixed by a previous design focused on determining if genetic variants coding for cytokines involved in inflammatory signaling were associated with silicosis (Yucesoy et al., 2001a,b). For this reason, the proportion of cases and consequent measures of association (odds ratios derived from logistic regression models) cannot be used to estimate population risk or relative risk. Only population-based
studies can be used to appropriately assess the demographics of silica-mediated fibrotic lung disease and the distributions of exposure and genetic determinants, as well as the attendant risks of disease. However, the relative features of the study design are sufficient to draw inferences about the role of genetic data for risk assessment and individual classification.

The method of categorization will affect the results and conclusions concerning the relative proportion of cases by risk decile. Categorization is always a balance between dose–response characterization and statistical precision. Too few categories may result in the inability to see an effect, while too many categories reduces the sample size and thereby the power to see an effect. If for instance, quantiles rather than deciles are used to categorize the relative proportions of cases, the observed difference between the two models (Fig. 1) would be masked by collapsing the first two deciles. For this reason, risk deciles, rather than other categorizations, are presented here to provide the greatest frequency of distinct categories while still maintaining sufficient sample sizes within each group to characterize dose–response features.

As demonstrated by the ROC curves, neither model was particularly effective as a diagnostic method for classification of individual disease status. The ideal ROC, representative of both high sensitivity and high specificity over a range of classification rules, would rise vertically from the left lower corner towards the left upper corner with little horizontal excursion, then rapidly change to predominantly horizontal excursion just below the horizontal grid line associated with 100% sensitivity. A poor curve, representative of only chance results, would be a diagonal straight line, running from the left lower corner to the right upper corner. The two curves plotted using our data fell between the two extremes, thus implying, first, that the two specified models were only marginally satisfactory for individual classification and, second, that the addition of genetic main effects and interactions does not effectively change the diagnostic ability of the model beyond occupational exposure alone.

It is this second implication that establishes the public health impact upon intervention strategies. That implication bears further scrutiny by examining the nature of measurement for occupational exposure, genetic polymorphisms and pathological assessment of disease. That nature includes issues of technical precision and validity of scientific constructs being examined.

The measure of occupational exposure was based solely on a report, by a second party, of years worked as an underground coal miner. Thus, the measure does not account for high or low exposure jobs, the fraction of particulates composed of silica, the fraction of respirable silica particulates, the effect of organic surface coating on a silica particle or recently fractured versus ‘old’ fractured silica particles, all of which have been implicated as causal mediators of silicosis (Green et al., 1989; Wallace et al., 1994; Castranova and Vallyathan, 2000). It is probable that the imprecision afforded by this very limited measure of occupational exposure notably underestimates relations between silica exposure and silicosis.

In contrast, the methods used to establish genetic polymorphic variants possess an extremely high degree of specificity, precluding the types of biases that may occur with life-time exposure measurements. Also, it is probable that these three polymorphisms capture the key genetic constructs related to these sources of variation associated with silicosis (Yucesoy et al., 2001a,b). This can also be said of the pathological assessment of silica-mediated lung fibrosis. This assessment is also technically specific, directly captures the key features of the disease and is not subject to the kind of significant misclassification that will bias measures of association.

Thus, the greatest, and probably only, source of misclassification error in this study is exposure, yet this simple measure of years worked as an underground coal miner is sufficient to adequately characterize heterogeneity of risk with pathologically assessed silica-mediated fibrosis of the lung. On the other hand, while the minor variants of IL-1\textalpha, IL-1RA and TNF\textalpha appear (i) to be highly prevalent, (ii) to play a mechanistic role in enhancing disease risk and (iii) to be associated with silicosis, they are not required for silicosis to occur. In contrast, exposure to silica (i) is a necessary determinant of silicosis, (ii) causes silicosis in a large proportion of miners who do not have minor variants in these three genes and, therefore, (iii) is always relevant to reduce the risk of disease. Therefore, the public health implications of an effective intervention strategy would appear to be limiting workplace exposure to protect workers against silicosis.

This conclusion, however, does not obviate the value or need to include genetics in research on occupational disease. The use of genetic research in epidemiological studies and in particular occupational disease studies is becoming more prominent, primarily because the role of the gene–environment interaction model is just as appropriate to the causes of toxicologically mediated disease as to pharmacologically mediated therapeutics. Genetic information in this study contributed to a better understanding of biological mechanisms in which products of these genes mediate and modulate deleterious outcomes leading to lung fibrosis. Moreover, this information, applied on a population basis, could be used to establish the most efficient public health preventive and therapeutic strategies to control not only lung fibrosis but a number of diseases involving common inflammatory processes.
Although the role of genetic information does not appear to play a necessary role as a causal determinant of occupationally mediated lung fibrosis and silica exposure does, this cannot be said for other occupationally mediated diseases. Indeed, genetic variants may be necessary determinants in the cause of immunologically mediated disease. Also, large differences in dose–response relations among genetic variants may suggest a value in using genetic information to prevent and treat toxicologically mediated disease. Thus, an understanding of biological mechanisms and of public health implications may well require joint assessment of gene and exposure information in occupational cohorts before rational and effective assessment and preventive and therapeutic strategies can be implemented. To this end, inclusion of genetics as a dimension of scientific research into the causes and determinants of occupational disease will be necessary. It is the ethical and moral use of such information that forms the basis for another level of inquiry, debate and intervention necessary to protect and aid the individual as well as society. It is a debate the scientific community can help frame by providing relevant and accurate understanding of the underlying science.

Acknowledgements—This study was supported by the Centers for Disease Control and Prevention at the National Institute for Occupational Safety and Health, Morgantown, WV.

REFERENCES


