USING CLASSIFICATION TREES TO IMPROVE CAUSAL INFERENCES IN OBSERVATIONAL STUDIES

Louis Anthony Cox, Jr.
Cox Associates and Cornerstone Consulting Group
503 Franklin Street, Denver, Colorado, 80218
TCoxDenver@aol.com

ABSTRACT

Much of the recent literature on AI and statistics has focused on how to use causal knowledge to enrich the set of valid causal inferences that can be drawn from available data and applied to practical problems such as decision-making, probabilistic diagnosis, and cost-effective control of systems. This paper examines classical problems of valid causal inference in observational studies, using epidemiological studies on the association between exposure to diesel exhaust (DE) and risk of lung cancer as a case study. It shows that one of the main applied computational tools of AI and statistics, classification tree analysis, can be adapted to help control or avoid many of the usual statistical threats to valid causal inference, and links this new use of classification trees to an established older literature on techniques for causal inference in social statistics based on elimination of competing (non-causal) explanations for observed associations. A strong link is then forged between an extension of classification tree analysis and modern AI and statistics approaches to causal modeling and inference based in directed acyclic graph (DAG) causal models and influence diagrams. This new link is based on the observation that classification tree analysis can be adapted to test the local Markov conditions that provide the critical defining structure of DAG models, as well as to quantify the conditional distributions of variables given the values of their parents – the key numerical information needed to quantify an influence diagram model. Finally, these insights are applied to available data on DE and lung cancer risks and are used to conclude that there is no evidence of a causal relation between them.

A. INTRODUCTION

In many areas of applied statistics, including epidemiology and social statistics, valid causal inferences are often threatened by artifacts of study design or data modeling that can create apparently statistically significant associations among variables in the absence of any corresponding causal relations. Notorious examples include model specification errors (e.g., from omitted heterogeneity in individual response functions, omitted measurement or classification errors in the independent variables, and omitted explanatory variables); selection artifacts in the study design or in data collection; and biases due to aggregation of data or to choice of data analysis methods. These threats can create statistically significant associations between the independent and dependent variables or lead to regression coefficients significantly different from zero in a generalized linear (GLIM) model, even if manipulating the independent variables would produce no corresponding changes in the dependent variables or would produce changes

opposite in sign from those indicated by the regression coefficients. Such artifacts can also mask true effects by biasing observed associations toward the null hypotheses of no relation.

As one of many examples, consider how statistically significant associations between exposure and response variables may be artificially created by the widespread practice of using estimated "relative risk" ratios (e.g., ratios of lifetime tumor probabilities, standardized mortality ratios for death-with-tumor, or ratios of age-specific hazard rates in exposed compared to unexposed populations) to report and compare exposure-associated risks in epidemiological risk assessments. If the statistical model is that each group (exposed and unexposed) has a certain risk rate that applies to each of its members, while the truth is that different individuals in each group have different risk rates, then reported relative risks may be consistently elevated (greater than 1) even if there is no difference in the frequency distributions of risk between the two populations. For example, suppose that X and Y are two independent, identically distributed random variables with the following probability density function: $Pr(X = 0.1) = 0.5$, $Pr(X = 0.4) = 0.5$ for X and $Pr(Y = 0.1) = 0.5$, $Pr(Y = 0.4) = 0.5$ for Y. In words, each variable is equally likely to have a value of 0.1 or 0.4. X might represent absolute risk (age-specific hazard rate or lifetime tumor probability) in an exposed population and Y might represent absolute risk in an unexposed population. Now, consider comparing risks based on the relative risk ratio X / Y. The average value of X / Y is $E(X/Y) = 0.25*(0.1/ 0.1) + 0.25*(0.1 / 0.4) + 0.25*(0.4 / 0.1) + 0.25*(0.4 /0.4) = (1 + 1/4 + 4 + 1) / 4 = 1.56$. Thus, the average value of the ratio X / Y is greater than 1. Randomly matching X individuals with Y individuals and comparing their values will lead to an average relative risk of about 1.56 even though there is no difference between the populations. This is an example of a bias due to the choice of modeling technique (namely, a relative risk ratio), which fails to adequately characterize inter-individual heterogeneity in risks. We shall refer to it as *ratio bias.*

A better-known but perhaps less prevalent source of bias due to omitted heterogeneity is illustrated by the 2 x 2 table in Table 1a.

TABLE 1A: HYPOTHETICAL AGGREGATE RISK DATA IN A 2 x 2 TABLE

	Exposed	Not Exposed
No Lung Cancer	595	901
Lung Cancer	505 (46%)	110 (11%)
	1100	1011

In this table, the relative risk of lung cancer among exposed workers compared to unexposed workers is RR = 0.46 / 0.11 = 4.2, which is statistically significantly greater than 1. Now suppose that this table resulted from aggregating two distinct occupational sub-populations, say with job codes A and B, having the descriptions in Tables 1b and 1c.

TABLE 1B: HYPOTHETICAL RISK DATA FOR JOB CODE A

	Job Code A (Heavily Exposed)	
	Exposed	Not Exposed
No Lung Cancer	500	1
Lung Cancer	500 (50%)	10 (91%)
	1000	11

TABLE 1C: HYPOTHETICAL RISK DATA FOR JOB CODE B

	Job Code B (Lightly Exposed)	
	Exposed	Not Exposed
No Lung Cancer	95	900
Lung Cancer	5 (5%)	100 (10%)
	100	1000

In the heavily exposed group A, the relative risk associated with exposure is 0.5 / 0.91 = 0.56, while in the less exposed group B, the relative risk is 0.05 / 0.10 = 0.50. Aggregation converts relative risks that are less than 1 for each group into an aggregate relative risk of 4.2 that is statistically significantly greater than 1 – a phenomenon known in the literature as Simpson's Paradox (Saari, 1987). The distinction between statistical significance and causal significance in this example is clear. The problem here is not with the choice of relative risk as a summary measure of association, but with the fact that aggregating data can create artificial patterns of association. Note that aggregation is logically equivalent to omission of a relevant explanatory variable (e.g., job code or extent of exposure) in modeling the association between exposure and response.

In summary, many of the most common methods for calculating and presenting statistical risk comparisons, including two-by-two tables and relative risk ratios, suffer from threats that undermine causal interpretation of the associations that they reveal.

This paper introduces an unexpected source of help in dealing with the uncertain relation between statistical associations and causal influences among variables. It discusses how *classification tree analysis,* one of the central tools emerging from the intersection of AI and statistics in the past dozen years (Biggs et al., 1991, Breiman et al., 1984; Buntine, 1993), can be adapted to control threats to valid causal inference in observational studies. Although they were originally intended primarily as a nonparametric, nonlinear alternative to traditional multivariate statistical methods for classification and regression, it turns out that classification trees can be adapted to address many of the most common threats to valid causal inferences.

B. MOTIVATION: DOES DIESEL EXHAUST CAUSE LUNG CANCER?

The basic idea for using classification trees to help improve causal inference builds on an established tradition in social statistics, which joins to the maxim "Correlation does not

imply causality" the qualification "Except when all other competing explanations can be ruled out" (Campbell and Stanley, 1963; Blalock, 1961). According to this tradition, statistical evidence can be used to establish the validity – or at least the likely usefulness – of an inference of causality precisely when it can be used to falsify all other hypotheses that might plausibly "explain away" the observed pattern of statistical association. These competing hypotheses reflect the types of threats and biases mentioned in the introduction. Indeed, a linchpin of the approach is the ability to systematically enumerate competing explanations (i.e., possible "threats" to a valid inference of causality) for an observed statistically significant association and then subject each explanation to statistical tests based on available data. This approach potentially allows causal inferences to be drawn even in "quasi-experiments" where direct experimental manipulation of independent variables is impossible – for example, because the data being analyzed were collected long ago under conditions that can no longer easily be replicated (Campbell and Stanley, 1963). This makes *inference based on elimination of competing explanations* a particularly valuable potential strategy for analyzing retrospective epidemiological studies. However, the cost of the strategy has traditionally been high: unless available data can rule out all (or all plausible) competing explanations, no firm conclusions about causality can be drawn. This may be seen as an inherent limitation on the ability to draw sound inferences about causation when direct experimental intervention is impossible.

To examine how classification trees can help improve causal inference based on elimination of competing explanations, we focus on two real data sets that have been used to draw causal inferences about the relation between exposure to diesel exhaust (DE) and lung cancer. The first is a "meta-data" set consisting of the reported results of 34 epidemiological studies (Cohen and Higgins, 1995; Muscat and Wynder, 1995). Populations examined included railroad workers in the United States and Canada, truck drivers in the U.S., bus garage workers and bus company employees in Sweden and in England, dock workers in Sweden, heavy equipment operators in the U.S., Minnesota highway maintenance workers, and the general population in the U.S. The second data set consists of detailed individual-level data on estimated occupational exposures, death times, and lung cancer mortality for over 55,000 railroad workers (Garshick et al., 1987; Crump et al., 1991.)

The meta-data set contains conflicting results. Three studies found relative risks of lung cancer significantly less than 1 ($p < 0.05$) for DE-exposed populations compared to unexposed populations, while 11 studies showed relative risks significantly greater than 1 ($p < 0.05$) for the exposed populations. The remaining studies are inconclusive, but 25 studies reported point estimates of relative risks greater than 1, while only 9 reported point estimates less than 1. Analysts interpreting these data for decision and policy makers have presented the following contrasting conclusions.

> "The available evidence suggests that occupational exposure to diesel exhaust from diverse sources increases the rate of lung cancer by 20% to 40% in

exposed workers generally and to a greater extent among workers with prolonged or intense exposure." (Cohen and Higgins, 1995)

"It can be concluded that short-term exposure to diesel engine exhaust (< 20 years) does not have a causative role in human lung cancer. There is statistical but not causal evidence that long-term exposure to diesel exhaust (> 20 years) increases the risk of lung cancer for locomotive engineers, brake men, and diesel engine mechanics [but] using common criteria for determining causal associations, the epidemiologic evidence is insufficient to establish diesel engine exhaust as a human lung carcinogen." (Muscat and Wynder, 1995)

Similarly, the more detailed data of Garshick et al. (1988) have been interpreted as showing a statistically significant positive association between DE exposure and lung cancer risk (Garshick et al., 1987; Cohen et al., 1995) and also, conversely, as showing a significantly negative association between DE exposure and lung cancer risk (Crump et al., 1991), depending on the statistical models and analysis techniques used. Thus, the area is ripe for clarification using new methods such as those from AI and statistics.

C. THREATS TO VALID INFERENCE IN PAST DIESEL-EXHAUST STUDIES

Reexamining the 34 epidemiological studies on DE exposure and lung cancer from the standpoint of quasi-experimental designs and valid causal inference (Campbell and Stanley, 1963) reveals several potential threats to validity that have not been adequately controlled in the study designs and analyses. This section summarizes the main ones. Its purpose is not to establish that any (or all) of these threats do explain away the apparent positive association between DE and lung cancer risk, but only to point out that they plausibly could – and that therefore, by the usual logic of proof-by-elimination, causality has not been established. This supports the second of the conflicting interpretations quoted above. Similarly, the following section, which indicates how classification tree analysis can be used to address many of the threats to valid causal inference, does not reanalyze each data set, since the raw data, spanning many countries and several decades, are not readily available. Instead, it only points out how classification trees could be used to avoid or ameliorate many of the threats to valid causal inference. However, the following section does apply classification trees to reanalyze the most influential of the 34 studies, namely that of Garshick et al. (1988), which has been used by policy makers and their advisors to support a conclusion that DE causes increased risk of lung cancer (Cohen and Higgins, 1995).

In the absence of a true causal relation, why might the majority of the epidemiological studies in the meta data base nonetheless find a positive association between DE and lung cancer risk? Table 2 lists potential *a priori* explanations that could rival the hypothesis of a causal relation.

TABLE 2: POTENTIAL THREATS TO VALID CAUSAL INFERENCES

1. *Multiple hypothesis testing bias.* (See discussion in text)
2. *Misspecified statistical models*, including those with the following limitations:
   - *Ignored estimation and classification errors* in estimated exposure histories. Such "errors-in-variables" can bias risk estimates toward the null by diluting a true association with irrelevant observations, or can overstate a small effect by attributing to each exposure level responses that are more likely to be due to mis-estimated higher exposures.
   - *Omitted heterogeneity,* meaning heterogeneity present in the data-generating process but not represented in the statistical model, in individual response probabilities. The "ratio bias" discussed in the introduction is an example.
   - *Omitted explanatory variables,* in which one or more factors not included in the model, but positively associated with both exposure and response, explain away the apparent positive exposure-response association.
   - *Wrong model form,* in which the set of candidate models considered for relating exposure to response tends to create a bias toward a positive association. See discussion in text.
3. *Biases due to data aggregation* (e.g., Simpson's Paradox) and data quantization
4. *Non-randomly missing and censored data* in which the censoring patterns is associated with either the dependent or the independent variables ("informative" censoring). *Example:* If cause-of-death information is more complete among exposed than among unexposed workers, then a relative over-counting of lung cancers could occur for the exposed group compared to the unexposed group.
5. *Other sampling and selection biases,* which cause the sample observations to be non-representative of the population from which they were sampled and for which inferences are to be drawn. Examples include artifacts in which the experience of being surveyed itself causes a change in reported job classifications, exposure histories, or reported incidence of lung cancers.

In the terminology of Campbell and Stanley (1963), who present a much more detailed taxonomy of possible omissions and selection artifacts that threaten valid causal inference, threats 4 and 5 affect the external validity of a study (do the conclusions drawn from the sample apply to the population of interest?), while threats 1 - 3 affect its internal validity (do the conclusions drawn hold even for the sample observations?)

The preceding threats are *a priori* logically possible hypotheses which – in addition to a true causal relation or random chance – might be advanced as potential explanations for an association between exposure and response in any epidemiological study. The ones that are relevant for a particular study are just those that have not been eliminated by the study design or adequately controlled for in the modeling and analysis of the study data. Reviewing the designs and analyses of the 34 studies on DE and lung cancer suggests that the following are the main threats that have not been adequately eliminated:

1. Multiple hypothesis testing bias. Many studies tested for a positive exposure-response relation in each of several subsets of the study population (e.g., in multiple age groups, exposure groups, and/or job categories) and then reported the "statistically significant" positive associations without reducing their p-values to compensate for the expected increase in false positives from testing multiple hypotheses on the same data. Clearly, this will tend to increase the expected number of false positives beyond what the reported p-values would allow.

> *Example:* Garshick et al. (1987, p. 1242) report that "Workers 64 years of age or younger at the time of death with work in a diesel exhaust exposed job for 20 years had a significantly increased relative odds (odds ratio = 1.41, 95% CI = 1.06, 1.88) of lung cancer." This is a special case of the statement template "Workers A years of age or younger exposed to diesel exhaust for at least Y years have increased relative odds ratios for lung cancer." The probability of a false positive occurring among the multiple hypotheses representing different age ranges (such as A =54, 59, 64, 69, etc.) and different durations of exposure (e.g., Y = 10 years, 15 years, 20 years, etc.) is not limited to p = 5% when this is the p-value applied to each pair of values. Therefore, the statement that the hypothesis is supported "with 95% confidence" when A and Y are instantiated as (A = 64, Y = 20) is erroneous: it amounts to selecting one subset of the population that favors the hypothesis and concluding that the hypothesis holds in general, without adequately testing whether the association is due to chance. Similarly, a subsequent cohort study (Garshick et al., 1988, p. 823) concluded that "We demonstrate an association between diesel exhaust exposure and lung cancer", although this conclusion is demonstrated only among workers with at least 15 years of exposure for whom exposure in the 4 years before death was disregarded – a data truncation that generated a positive result in this study but *not* in the case-control study, where "the relative odds ratio of lung cancer decreased slightly with recent exposure disregarded" (*ibid.*, p. 823). A positive result created only by selectively discarding data cannot correctly be considered "statistically significant" unless the probability is considered of achieving such a result by chance when all possible data truncations are allowed.

2. Misspecified statistical models. Many regression and generalized linear models used in risk assessment are unable to show non-monotonic (or negative) relations between exposure and response variables. Using such models can artificially create positive associations where a less constraining model would not. As previously mentioned, statistical risk models may also be misspecified by omitting errors in exposure estimates from the model formula; neglecting heterogeneity in individual response probabilities; and ignoring explanatory variables that affect response probabilities.

> Example: The influential Garshick et al. cohort studies previously cited used two statistical risk models – proportional hazards and conditional logistic regression – which assume that cumulative exposure affects risk only through a quantity bx, where x denotes cumulative exposure and b is a cancer potency parameter for DE.

In the current application, these models suffer from at least the following misspecification errors:

- *Uncertainty about the correct value of DE exposure* for each individual is not represented in the model.

- *Inter-individual heterogeneity and variability* in responses are ignored. Both models implicitly assume that the value of b is the same for all individuals in the same exposure group. More generally, all of the 34 studies examined reported estimated risk ratios but did not model inter-individual heterogeneity in response probabilities, thus leaving open the possibility that the ratio bias identified in the introduction explains the apparent positive associations between DE and lung tumors.

- *Wrong model form.* Both models assume that the value of b is *independent of exposure* and that it is *constant over time.* In addition, both the proportional hazards model and the conditional logistic regression model formulas imply that risk is proportional to exposure at all sufficiently small exposure levels (since both approach straight-line relations for small argument values). This specification is inconsistent with what is known about the biology of lung cancer and with recent evidence on nonlinearities and threshold-like responses in DE-related lung tumors found in animals (e.g., Driscoll et al., 1996). Biological knowledge indicates that DE lung cancer potency depends strongly on prior exposure history and increases sharply with dose concentration and duration above a certain threshold (Henderson et al., 1988; Driscoll, 1996).

In summary, as suggested by these examples, at least threats 1 and 2 in Table 2 are relevant to the past epidemiological studies of lung cancer causation by DE. (For particular studies, threats 3- 5 are also relevant.) One reaction to this state of affairs could be to simply dismiss the accumulated epidemiological evidence as inconclusive, due to limitations in study designs and analyses that fail to eliminate the threats and biases in Table 2 as plausible explanations for observed associations. A more constructive response is to invoke new techniques to determine whether a causal relation can be definitely established (by ruling out other plausible explanations in retrospect) or definitely refuted (by exhibiting a specific alternative explanation) using the data from key past studies. This approach is developed next.

D. SUPPORTING VALID CAUSAL INFERENCE WITH CLASSIFICATION TREES

Recall that classification tree algorithms are based on the key idea *of recursive partitioning,* in which the population of interest is successively partitioned into increasingly homogeneous sub-groups. Homogeneity (or "purity") is measured by a criterion such as the average classification entropy of the conditional distributions of the response variable (Breiman et al., 1984; Buntine, 1993), deviance (Venables and Ripley,

1994) or by a chi-square or F-test (Biggs et al., 1991) for homogeneity of the conditional distribution of responses. Each successive partitioning ("branching") is accomplished by conditioning on the level of one independent variable (a "node"). Such algorithms create "classification trees" in which the internal nodes (also called "splits") represent conditioning variables and their branches represent values (or ranges of values, for continuous variables). A path through such a tree represents a conjunction of independent variable values that lead to a relatively homogeneous conditional distribution for the value of the dependent variable. The leaf nodes represent conditional distributions that cannot be further resolved using the selected criterion for measuring homogeneity.

Classification tree algorithms provide a nonparametric alternative to other multivariate classification and regression methods that tends to be especially well suited for detecting and exploiting strong, nonlinear interactions among independent variables in determining the value of the dependent variable. They can be used for categorical, ordered categorical, continuous, and mixed data types, and make no specific hypotheses about the form of the multivariate relation between the independent and dependent variables. These characteristics make classification tree algorithms appropriate, with only minor changes, for investigating and controlling for many of the threats to valid causal inference listed in Table 2. Specifically, the most important threats affecting inferences about causation for the DE-lung tumor association can be controlled by the following techniques.

Control of multiple hypothesis testing bias. Classification tree algorithms deliberately investigate many subsets of individuals, corresponding to different paths through the tree, and attempt to construct subsets (identified by specified combinations of factor values) that exhibit relatively homogeneous, and hence predictable, responses. Thus, hypotheses-testing about many subsets of the population in parallel is an explicit part of the methodology. Moreover, commercial implementations such as KnowledgeSeeker™ (Biggs et al., 1991) automatically group categories (respecting ordinal constraints if the categories are ordered) and select intervals for continuous variables in an effort to maximize predictability of the dependent variable. Hence, there is an intrinsic need to correct for multiple hypothesis-testing and for the fact that general statement-templates (e.g., "The value of factor X is at least x") are being instantiated before the statistical significance of a split can be correctly assessed. Several techniques are therefore built into current classification tree algorithms to adjust the p-values appropriately for specific subsets and splits. For example, KnowledgeSeeker™ uses Bonferroni inequalities to reduce p-values of splits to adjust for the fact that the "best" (most predictive) grouping of categories or selections of intervals are used. The performance of this relatively simple approach has been validated by Monte-Carlo simulation (Biggs et al., 1991). CART (Breiman et al., 1984) uses a more computationally intensive approach based on cross-validation and approximately optimal "pruning" of trees to compensate for over-training on specific data, while many modern algorithms, including recent releases of KnowledgeSeeker™, use estimated re-substitution error rates to quantify and control for false positives due to multiple hypothesis testing on a single training data set. Because

the need to control for biases due to multiple hypothesis testing has been well recognized and built into these algorithms, details of the techniques may be found in the references.

Avoiding biases due to model misspecifications. Model specification errors can be reduced or eliminated in classification tree models by avoiding the necessity of specifying an assumed mathematical formula to relate independent and dependent variables (e.g., a specific linkage function for a GLIM model). As in influence diagrams, dependencies in classification trees are quantified by *conditional distributions* of the values of some variables – specifically, the dependent variable – given the values of its predecessors (the values of the independent variables along branches leading to the leaf nodes at which the conditional distributions of the dependent variables are assessed.) Conditional distributions provide a rich, flexible language for expressing dependencies without assuming constraining parametric forms. Hence, thresholds and non-monotonic relations, which may not be detected or allowed by GLIM and proportional hazards models, can be revealed clearly by classification trees. Thus, classification trees can avoid some biases due to "wrong model form" specification errors that may be important for analyzing DE-lung cancer data sets. (An open research question is what kinds of common multivariate relations, if any, cannot be well expressed using conditional distributions. One example is that simple linear relations cannot be concisely represented using trees such as those in KnowledgeSeeker™ that rely on piecewise-constant approximations; on the other hand, they can be expressed concisely in a CART tree, which allows for piecewise linear relations among independent and dependent variables at its leaf nodes.)

Other potential biases due to model specification errors are at least partly addressed by classification tree methodology. For example, it is well known that omitted measurement errors in independent variables can attenuate their estimated coefficients in linear regression models, leading to false negatives when the hypothesis is tested that their coefficients differ significantly from zero (Blalock, 1961). Classification trees replace the often intractable problem of testing hypotheses about coefficients of unobserved variables (e.g., of true exposure when only estimated exposures are measured) with the simpler hypothesis-testing problem of testing whether conditional distributions of response variables differ significantly for different values of the measured independent variables (e.g., estimated exposures). By focusing entirely on conditional relations among empirically measured quantities, classification trees avoid some of the problems associated with "true but unmeasured" constructs (Blalock, 1961). The cost may be a less theoretically rich model. For example, a classification tree model may be silent about the relation between true but unmeasured DE exposure and lung cancer risk.

Finally, biases due to the other sources in Table 2 – specifically, data aggregation, ignored heterogeneity in individual responses, and omitted explanatory variables – can be partly addressed by modifying the usual classification tree algorithms to test for *latent splits* at leaf nodes. Even though a leaf node in a fully developed classification tree is constructed so that conditioning on any of the dependent variables will not produce a further statistically significant split (as defined by the particular tree-growing algorithm used), it is possible that some other variable, not included among the identified independent

variables, explains some of the remaining variance in the response variable. It is surprisingly often possible to test the hypothesis that such a "latent" independent variable, meaning one that is not included in the model but that significantly affects the dependent variable, exists, even though its identity clearly cannot be determined by statistical analysis alone. As an example, consider a dichotomous response variable such as "dead with lung tumor" vs. "not dead with lung tumor". (For extensions of classification trees to deal with censored survival data such as time-to-tumor data, see Bacchetti and Segal, 1995). If a classification tree successfully sorted individuals into homogeneous groups based on the identified independent variables, then all individuals assigned to the same leaf node would have the same response probability, which could be estimated as the sample proportion of positive responses at that leaf node. In this case, the variance of the individual responses would (perhaps counter-intuitively) be larger than if different individuals had different response probabilities (Feller, 1968). Thus, the variance around the mean can potentially be used to determine whether all response probabilities are equal, so that true homogeneity has been achieved. More generally, a mixture of up to k distinct binomial distributions can in principle be identified at a leaf node with N members if and only if $k \leq (N + 1) / 2$ (Titterington et al., 1985, p. 40.) Each such component of a mixture distribution detected by analysis of the frequency distribution of responses at a leaf node may be represented as a further branch. The former leaf node may then be referred to as a *latent split,* since the k new leaf nodes descending from it have been inferred via mixture distribution analysis and may be interpreted as being conditioned on the k different possible values of a "latent" dependent variable that has not been explicitly included in the model. Latent splits, which can also be developed for multinomial and continuous multivariate distributions under certain identifiability conditions (Titterington et al., 1985), provide one constructive approach to the detection of omitted explanatory variables and modeling of residual (unexplained) heterogeneity in the conditional response distributions at leaf nodes.

The previous paragraphs have focused on data analysis and modeling approaches to overcoming threats 1-3 in Table 2. Study designs and analyses for "quasi-experiments" that address threats 4 and 5 are discussed in detail by Campbell and Stanley (1964). Classification trees can make one other major contribution to this style of causal analysis, using a modified version of surrogate splits (Breiman et al., 1984; Venables and Ripley, 1994). It is usually the case that an internal node ("split") in a classification tree will be the top-ranked member (according to the criterion selected for the recursive partitioning algorithm) of a set of several possible independent variables, any of which could potentially serve as the node on which a split is performed. As Breiman et al. (1984) pointed out, this fact can be exploited to bypass missing data by substituting an alternative "surrogate split" for the top-ranked split when data are not available to implement the top-ranked split. This idea can be extended to other reasons for choosing an alternative to the top-ranked split. For example, if the goal is changed from pure statistical inference to cost-effective control of the dependent variable, then the top-ranked variable may be rejected if it is not controllable by the decision maker, or if it is too expensive to measure or to change. (Alternatively the split criterion can be modified to take these other objectives into account.) If the goal is to test a causal hypothesis, then

the split-selection criterion may be modified to give priority to variables other than the postulated causal predecessor(s) of interest (DE exposure variables, in this case) to determine whether combinations of other variables suffice to "explain away" the heterogeneity in the response variable.

A general use of classification trees in causal modeling is as follows, assuming that the structure of the underlying causal model is to be represented by a directed acyclic graph (DAG) model (e.g., Yao and Tritchler, 1996). Suppose that a set of N variables is partitioned into a single dependent variable of interest, say j; a set of immediate predecessors (i.e., parents) of j, denoted by P(j); a set of all j's immediate successors and their descendants, D(j), and a set of other variables, say N(j), that are disjoint from j, P(j), and D(j). If the partial structure defined by {j, P(j), D(j), N(j)} is correct, then a classification tree grown for j using only the variables in P(j) should have the property that it is stable when variables in N(j) are allowed as candidate splits, meaning that allowing additional tree-growing using variables in N(j) following construction of an initial tree using only the members of P(j) should not result in any additions to the tree from N(j). In other words, the value of j should be conditionally independent (CI) of the values of variables in N(j), given the values of variables in P(j) – a key concept sometimes referred to as the "local Markov condition" in influence diagrams and DAG causal models (Yao and Tritchler, 1996). Thus, classification tree algorithms can be used to check whether the local Markov conditions in an assumed DAG causal model hold. Extending this reasoning, classification trees also provide a constructive way to use data to quantify the conditional probability distribution for j, given combinations of values of variables in P(j) (corresponding to paths through the tree when j is taken the dependent variable and the tree is grown using variables in P(j)); and they can support systematic search for the members of P(j) as the smallest set of variables such that j is CI of variables in N(j) given j. (Technically, to distinguish $X \rightarrow Y \rightarrow Z$ from $Y \rightarrow X \rightarrow Z$, it is necessary to assume that X and Y are not deterministic one-to-one functions of each other, so that the first chain implies that Z is CI of X given Y, but not that Z is CI of Y given X.) In summary, classification tree algorithms provide valuable tools for applying multivariate data to help construct, quantify, and validate conventional DAG causal models, since they can be used to test the key CI relation or local Markov condition in such models can by tree growing with constrained splits. This capability also allows classification trees to refute erroneous causal models by showing that the postulated parents of j do not shield j from variables modeled as being in N(j), i.e., by exhibiting violations of the local Markov conditions implied by a hypothesized causal model.

E. REANALYSIS OF CAUSATION IN AN EPIDEMIOLOGICAL DATA SET

The preceding ideas were applied in a classification tree analysis of the detailed individual data of Garshick et al. (1988), with updated exposure estimates by Crump et al. (1991), for railroad workers occupationally exposed to DE. The main findings were as follows.

- Several strong interrelationships among variables were detected by growing classification trees for different choices of dependent variables. Specifically, worker age in 1959 (when dieselization of the train engine fleet was essentially complete) was strongly positively associated with estimated average exposure concentration during exposure, but significantly negatively associated with the total number of months of exposure. Similarly, estimated average exposure concentration was significantly positively associated with age at death. These inter-correlations among independent variables suggest why regression models of the partial effects of exposure concentration or exposure duration on lung cancer risk must be interpreted very carefully, and why their signs may change depending on exactly which variables are included in the regression.

- Incidences of Simpson's paradox were discovered. For example, age in 1959 and age at death are significantly positively associated overall, and yet the relation is negative within each group of workers exposed to the same average concentration level of DE. Thus, a set of negative relations (stating that workers who were younger in 1959 tend to live longer) aggregate to a positive relation.

- The association between worker age in 1959 and probability of dying with lung cancer is non-monotonic: it is significantly positive among workers dying before 1966, but significantly negative among workers dying after 1974. This nonmonotonic pattern was easily detected by the classification tree analysis, but clearly implies that no single regression coefficient (e.g., in a logistic regression or other GLIM model) can adequately summarize the relation between age in 1959 and lung cancer mortality risk. Similarly, the relation between average exposure concentration and risk of lung cancer was nonmonotonic within specific age groups. These observations justify in retrospect the theoretical concern raised in Table 2 about possible model misspecifications in previous analyses.

- Cumulative months of DE exposure is strongly positively associated with lung tumor risk (and also with total duration of employment between 1959 and retirement), but estimated average concentration of DE exposure during the months of exposure seems not to be positively associated (and indeed, based on a supplementary survival data analysis, appears to be significantly negatively associated) with lung tumor risk. This suggests that some other factor in the occupational environment, rather than DE itself, may be causing increased lung tumor risks.

- Prediction of lung tumor mortality rate can be based on either of the following two sets of independent variables: P1 = {months of exposure, age at death} or P2 = {year of death, age at death, duration of employment between 1959 and retirement}. Both are stable sets, in the terminology of the previous section: given the values of either set of variables, no other variables (including estimated average DE exposure concentrations) improves ability to predict lung tumor mortality using any of the tree-growing options provided by the KnowledgeSeeker™ classification tree analysis package of Biggs et al. (1991).

In summary, after using classification tree analysis to correct for multiple hypothesis testing biases and model specification errors (wrong model form) that were found in previous analyses of this data set, no significant association between DE exposure and lung tumor mortality risk remained. To the contrary, the predictive power of the data set to predict lung cancer risk appears to be exhausted by the set of variables P2 = {year of death, age at death, duration of employment between 1959 and retirement}, with no further contribution being made by exposure-related variables such as the months or estimated average concentration of DE exposure once these other variables have been accounted for. These findings demonstrate that there no evidence of a causal association between DE exposure and lung cancer morality risk remains after conditioning on the values of other variables. Since the variables in P2 are strongly (but not always monotonically) related to each other and to the exposure variables, however, it is not surprising that there are statistical associations among the exposure variables and lung cancer mortality risks. The sign of the association depends on the values of other variables, in contrast to the usual assumption made in previous multiple linear regression and GLIM analyses of this data set (Garshick et al., 1988; Crump et al., 1991).

F. CONCLUSIONS

The modified classification tree analysis discussed in this paper provides a potentially powerful applied tool for using multivariate data to test causal theories and hypotheses (e.g., the hypothesized structures of DAG causal models); for estimating conditional probability distributions of nodes in influence diagrams; and for controlling for sources of bias and for common threats to valid causal inference in observational studies (specifically, threats 1 and 2 in Table 2, and perhaps threat 3 when the proposed device of "latent splits" can be used). These points have been developed in the context of epidemiological analyses of the possibility of a causal relation between exposure to diesel exhaust (DE) and subsequent lung cancer mortality risk. The classification tree approach shows that, in contrast to interpretations of some previous statistical analyses, available data and studies do not justify an inference of a causal relation between DE exposure and lung cancer mortality risk. Instead, the available evidence is consistent with the existence of artifactual statistical associations between DE exposure and lung cancer mortality risks in the absence of a causal relation. Although this is a negative conclusion, the value of the classification tree analysis technique has been clearly demonstrated in revealing non-monotonic relations among variables (violating the assumptions of previously applied statistical model) and in identifying a set of non-exposure variables that fully exhaust the predictive power of the data set for predicting lung cancer mortality risks – without positing any relation to DE exposure. An intriguing suggestion from the case study data analysis is that there may be a factor other than DE in the occupational environment that contributes to increased risk of lung cancer mortality. This unknown factor could be treated as a possible "latent split" variable in future classification tree analyses.

REFERENCES

Bacchetti, P., and M.R. Segal, 1995. Survival trees with time-dependent covariates: Application to estimating changes in the incubation period of AIDS. *Lifetime Data Analysis,* 1, 1, 35-48.

Biggs, D., B. de Ville, E. Suen, 1991. A method of choosing multiway partitions for classification and decision trees. *J. Applied Statistics,* 18, 1, 49-62.

Buntine, W., 1993. Learning classification trees. In D.J. Hand (Ed.), *Artificial Intelligence Frontiers in Statistics: Artificial Intelligence and Statistics IV.* Chapman andd Hall, 182-201.

Blalock, H.M., 1961. *Causal Inferences in Nonexperimental Research.* University of North Carolina Press, Chapel Hill.

Campbell, D.T., and J.C. Stanley, 1963. *Experimental and Quasi-Experimental Designs for Research.* Rand McNally, Chicago.

Cohen, A.J., and M.W.P. Higgins, 1995. Health effects of diesel exhaust: Epidemiology. In Diesel Exhaust: A Critical Analysis of Emissions, Exposure, and Health effects. A Special report of the Institute's Working Group. Health effects Institute (HEI), Cambridge, MA.

Crump, K.S., T. Lambert, and C. Chen, 1991. Assessment of risk from exposure to diesel engine emissions. Report prepared for the U.S. Environmental Protection Agency, Office of Health Assessment, by Clement International Corporation, Ruston, Louisiana (Work Assignment No. 182, July).

Driscoll, K.E., J.M. Carter, B.W. Howard, D.G. Hassenbein, W. Pepelko, R.B. Baggs, and G. Oberdorster, 1996. Pulmonary inflammatory, chemokine, and mutagenic responses in rats after subchronic inhalation of carbon black. *Toxicology and Applied Pharmacology,* 136, 372-380.

Feller, W., 1968. *An Introduction to Probability Theory and Its Applications, Volume 1, Third Edition.* Wiley, New York, p. 231.

Garshick, E., M.B. Schenker, A. Munoz, M. Segal, T.J. Smith, S.R. Woskie, S.K. Hammond, F.E. Speizer, 1988. A case-control study of lung cancer and diesel exhaust exposure in railroad workers. *American Review of Respiratory Diseases,* 135, 1242-1248.

Garshick, E., M.B. Schenker, A. Munoz, M. Segal, T.J. Smith, S.R. Woskie, S.K. Hammond, F.E. Speizer, 1988. A retrospective cohort study of lung cancer and diesel exhaust exposure in railroad workers. *American Review of Respiratory Diseases,* 137, 820-825.

Henderson, R.F., J.A. Pickrell, R.K. Jones, J.D. Sun, J.M. Benson, J.L. Mauderly, R.O. McClellan, 1988. Response of rodents to inhaled diluted diesel exhaust: Biochemical and cytological changes in bronchoalveolar lavage fluid and in lung tissue. *Fundamental and Applied Toxicology,* 11:546-567.

Judge, G.C., W.E. Griffiths, R.C. Hill, H. Lutkepohl, and T-C Lee, 1985. *The Theory and Practice of Econometrics: Second Edition.* Wiley, New York.

Kalbfleisch, J.D., and R.L. Prentice, 1980. *The Statistical Analysis of Failure Time Data.* Wiley, New York.

Mauderly, J.L., R.K. Jones, W.C. Griffith, R.F. Henderson, and R.O. McClellan, 1987. Diesel exhaust is a pulmonary carcinogen in rats exposed chronically by inhalation. *Fundamental and Applied Toxicology,* 9:208-221.

McCullagh, P., and J.A. Nelder, 1983. *Generalized Linear Models.* Chapman and Hall, New York.

Muscat, J.E. and E.L. Wynder, 1995. Diesel engine exhaust and lung cancer: An unproven association. *Environmental Health Perspectives,* 103(9), 812-818.

Nakamura, T., 1992. Proportional hazards model with covariates subject to measurement error. *Biometrics* **48**, 829-638.

Pearl, J., 1995. Causal diagrams for experimental research (with discussion). *Biometrika,* **82**, 4, 669 - 709.

Saari, D.G., 1987. The sources of some paradoxes from social choice theory and probability. Journal of Economic Theory, **41**, 1-22.

Spirtes, P., C. Glymour, and R. Scheines, 1993. Causation, Prediction, and Search. Springer-Verlag, New York.

Stober, W., and J.L. Mauderly, 1994. Model-inferred hypothesis of a critical dose for overload tumor induction by diesel soot and carbon black. *Inhalation Toxicology,* **6**:427-457.

Titterington, D.M., A.F.M. Smith, and U.E. Makov, 1985. *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York.

Venables, W.N., and B.D. Ripley, 1994. *Modern Applied Statistics with S-Plus.* Springer-Verlag, New York.

Yao, Q., and D. Tritchler, 1996. Likelihood-based causal inference. Chapter 4 in . Fsher and HJ Lenz (eds). *Learning from Data: Artificial Intelligence and Statistics V.* Springer Verlag, New York, 1996.