Published in final edited form as:
Occup Environ Med. 2013 March ; 70(3): 203–210. doi:10.1136/oemed-2012-100918.

Inside the black box: starting to uncover the underlying decision rules used in one-by-one expert assessment of occupational exposure in case-control studies

David C. Wheeler[1,2], **Igor Burstyn**[3], **Roel Vermeulen**[4], **Kai Yu**[5], **Susan M. Shortreed**[6], **Anjoeka Pronk**[7], **Patricia A. Stewart**[8], **Joanne S. Colt**[1], **Dalsu Baris**[1], **Margaret R. Karagas**[9], **Molly Schwenn**[10], **Alison Johnson**[11], **Debra T. Silverman**[1], and **Melissa C. Friesen**[1]

[1]Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda MD [2]Now at: Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA [3]Drexel University, Philadelphia, Pennsylvania [4]Utrecht University, Utrecht, Netherlands [5]Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda MD [6]Biostatistics Unit, Group Health Research Institute, Seattle, Washington [7]TNO, Zeist, Netherlands [8]Stewart Exposure Assessments, LLC, Arlington, VA [9]Dartmouth Medical School, Hanover, New Hampshire [10]Maine Cancer Registry, Augusta, Maine [11]Vermont Cancer Registry, Burlington, Vermont

Abstract

Objectives—Evaluating occupational exposures in population-based case-control studies often requires exposure assessors to review each study participants' reported occupational information job-by-job to derive exposure estimates. Although such assessments likely have underlying decision rules, they usually lack transparency, are time-consuming and have uncertain reliability and validity. We aimed to identify the underlying rules to enable documentation, review, and future use of these expert-based exposure decisions.

Methods—Classification and regression trees (CART, predictions from a single tree) and random forests (predictions from many trees) were used to identify the underlying rules from the questionnaire responses and an expert's exposure assignments for occupational diesel exhaust exposure for several metrics: binary exposure probability and ordinal exposure probability, intensity, and frequency. Data were split into training (n=10,488 jobs), testing (n=2,247), and validation (n=2,248) data sets.

Corresponding Author: Melissa C. Friesen, Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd, Room 8106 MSC 7240, Bethesda MD 20892-7240 Telephone: (301) 594-7485; Fax: (301) 402-1819; friesenmc@mail.nih.gov.

Contributors DCW and MCF designed the statistical learning analysis approach to assess the participants' exposure to diesel exhaust. DCW conducted all statistical analyses. IB, RV, KY, and SMS assisted in the statistical design. DCW, MCF, IB, RV, KY, SMS, AP, PAS, and DTS provided interpretation of the methods and their application. PAS, JSC, DB, MRK, MS, AJ, SC and DTS initiated and designed the bladder cancer case-control study, including the development of tools to collect occupational information, and supervised all aspects of data collection and uses of the study data. DCW and MCF drafted and revised the paper based on feedback provided from all authors.

Competing interest None.

Results—The CART and random forest models' predictions agreed with 92–94% of the expert's binary probability assignments. For ordinal probability, intensity, and frequency metrics, the two models extracted decision rules more successfully for unexposed and highly exposed jobs (86–90% and 57–85%, respectively) than for low or medium exposed jobs (7–71%).

Conclusions—CART and random forest models extracted decision rules and accurately predicted an expert's exposure decisions for the majority of jobs and identified questionnaire response patterns that would require further expert review if the rules were applied to other jobs in the same or different study. This approach makes the exposure assessment process in case-control studies more transparent and creates a mechanism to efficiently replicate exposure decisions in future studies.

Keywords

diesel exhaust; classification; data mining; occupational exposure

INTRODUCTION

Exposure assessment of occupational risk factors in population-based studies is challenging. These studies rely on subject-reported lifetime occupational histories, and in some studies, on the subjects' responses to more detailed questions in job- or industry-specific modules. Typically, one or more exposure assessors review the questionnaire responses one job at a time to ascertain exposure – a time-consuming activity when each subject reports an average of 6 jobs over a lifetime [1–3]. While experts may document their decision rules, these exposure decisions are rarely explicitly published and thus provide no mechanism for others to evaluate or reproduce these assessments. This lack of transparency faces substantial criticism [4, 5]. As a result, alternative approaches are being implemented that use structured model-based exposure assessments to apply expert-based decision rules based on patterns in questionnaire responses [1, 6].

If decision rules can be successfully applied to questionnaire responses in epidemiologic studies, it raises the question of whether we can learn from the patterns of exposure decisions previously made by experts and apply them to other studies. We refer to these patterns as latent or underlying decision rules, because although the experts have used rules in making their assessments, the explicit rules that relate the questionnaire response patterns to the exposure decision may not have been documented. These underlying decision rules are valuable, given the time-consuming nature of the assessments and the limited number of available experts with broad knowledge about historical occupational situations.

To determine if underlying decision rules can be identified, we applied two statistical learning approaches designed to extract patterns and relationships between variables [7–8] — classification and regression trees (CART) and random forests—to questionnaire responses and the associated expert-based exposure estimates for occupational diesel exhaust exposure in the population-based New England Bladder Cancer case-control study [9]. Uncovering these rules has several important benefits. First, it can provide a mechanism for replicating the decision rules for other subjects within or across studies. Second, it can reduce the burden on the expert if the existing rules can be applied to the questionnaire responses and then reviewed by an expert, as was shown in the development of an asthma-specific job-exposure matrix [10]. Third, it makes the decision rules transparent, thus providing a way for other experts to evaluate and improve the rules.

METHODS

Model overview

We focused on tree-based statistical learning approaches because decision trees are able to handle non-linearity, interactions, and missing values, while producing interpretable decision rules [11–13]. Tree-based approaches predict decisions based on a sequential splitting pattern that resembles an upside-down tree, with the `root' at the top, below which are nodes that divide observations into branches. At the bottom are `leaves' that provide the predicted assignment (Figure 1). The nodes are selected iteratively, with the most predictive variable at each node used to split the observations into two branches according to that variable. Within each branch, the splitting continues until the model meets specified stopping criteria, such as a complexity parameter set to control the growth of the tree or a minimum number of observations per leaf. To illustrate, we show a fictional, simple decision tree in Figure 1 to classify 100 jobs into unexposed and exposed to diesel exhaust. The decision rules represented by the tree are revealed by starting at the root and evaluating the condition at each node to determine which branch to follow until a leaf (an exposure decision) is encountered (0 = unexposed; 1 = exposed, 1st label from top). Each leaf also reports the number of jobs assigned to the leaf (2nd label) and the proportion of jobs accurately classified within the leaf (3rd label). This tree has four leaves, a depth of two nodes, and uses three variables to classify jobs as exposed or unexposed. In this example, the most predictive variable is `smelled exhaust'. If the subject neither smelled exhaust nor worked on a construction site on a particular job, the model classified the job as unexposed (0), which agreed with the expert's assignments for 90% of the 10 jobs in that leaf. Conversely, if a subject smelled exhaust and was a truck driver, the model classified the job as exposed (1), with 90% agreement with the expert's assignments for the 30 jobs in that leaf.

Each decision tree can also be written as a series of conditions that provides a clear interpretation of the questionnaire response patterns that lead to an exposure assessment decision. For example, the conditions that created the second leaf from the left in Figure 1 are:

Rule: [exposure classification = 1; probability job assigned exposure by expert = 80%]

Smelled exhaust = no

Construction site = yes

The probability in the first line of the decision rule above is the relative frequency of the classified jobs that agreed with the expert assignment of being exposed (exposure classification = 1) and not smelling exhaust, but working at a construction site (32/40 in this case).

Random forest models are based on CART but can improve the predictive performance of CART by averaging the predictions across many simple trees [14]. In CART, the entire training data set and all entered variables are evaluated to derive one tree. In contrast, random forest models develop hundreds of trees, where each tree is trained on a random subset of the observations (jobs) and on a random subset of the input (questionnaire response) variables [15].

Study population

To examine whether or not CART and random forest models could identify an expert's decision rules, we used 14,983 jobs reported by the subjects from the New England Bladder

Cancer Study (n=1,213 cases, 1,418 controls) [9]. Each participant completed a lifetime occupational history questionnaire. The occupational history had open-ended questions asking the job title, name and location of employer, type of service or product provided, year started and stopped, work frequency, activities and tasks, the tools and equipment, and chemicals and materials handled. In addition, `did you ever work near diesel engines or other types of engines' and `did you ever smell diesel exhaust or other types of engine exhaust' were asked for each job. Answers to the occupational histories may have triggered a module that asked more detailed diesel exhaust and non-diesel exhaust questions for 67 jobs/industries; a module was completed for 64% of the reported jobs.

Diesel exhaust exposure estimates

The jobs were reviewed one-by-one by an industrial hygienist to assign the probability, intensity, and frequency of diesel exhaust exposure [6]. Probability was assessed based on the estimated proportion of workers likely exposed to diesel exhaust for the reported information including task, job or industry, and decade, with estimated cut points of <5% (none/negligible, category 0), 5–49% (low, 1), 50–79% (medium, 2), and ≥80% (high, 3). Approximately 75% of the jobs were assessed as having none or negligible probability of exposure. Intensity was assessed on a continuous scale as the estimated average level of respirable elemental carbon (REC, $\mu g/m^3$) in the workers breathing zone during tasks where diesel exhaust occurred and categorized with cut points of <0.25 (none/incidental, category 0), 0.25 to <5 (low, 1), 5 to <20 (medium, 2), and ≥20 (high, 3) $\mu g/m^3$ REC. Frequency was assessed on a continuous scale as the estimated average number of hours per week exposed to diesel exhaust and categorized with cut points of <0.25 (none/negligible, category 0), 0.25 to <8 hours per week (low, 1), 8 to <20 hours per week (medium, 2), and ≥20 (high, 3) hours per week.

Identifying questions related to diesel exhaust

We reviewed the responses to the occupational histories and the job- and industry-specific modules to identify variables that could be potential determinants of an expert's exposure assignment. All categorical variables were re-coded into dichotomous variables. This recoding does not change the information provided to the expert, it merely changes the form of the variables to a more convenient modeling structure. From the free-text responses in the occupational histories, we coded diesel exhaust information into standardized variables [6], resulting in 51 dichotomous occupational history variables, such as `job had traffic exposure', `job used diesel equipment', and `job start year'. We included variables for 83 2-digit and 169 3-digit standardized industry codes [16], and 61 2-digit and 134 3-digit standardized occupation codes [17]. From the module responses, we coded 67 dichotomous variables identifying the administered module, one variable indicating no module was completed, one variable indicating a module with diesel exhaust-related questions was completed, and 154 variables derived from questions directly or indirectly related to diesel exhaust exposure. Examples of module variables included `traffic-exposed job', `equipment powered by diesel', and `industry = heavy construction'. Overall, 498 variables were extracted from the occupational histories and 223 variables from the modules were extracted.

Model development

We used the rattle package [18,19] in R [20], which interfaces with the rpart [21] and randomForest [22] R packages to develop CART and random forests, respectively. Both approaches were used to predict a binary probability metric (none/low=0 vs. medium/high=1) to evaluate the models' ability to separate the jobs into exposed and unexposed categories, so that, at a minimum, the model predictions could focus the expert review on the more likely exposed jobs. The models were also used to predict ordinal metrics (0–3) for

probability, intensity, and frequency of exposure, which were treated as discrete, non-ordered categories in the model.

We randomly split the jobs into three datasets to get unbiased estimates for prediction errors for each model: 1) a training data set comprising 70% of the data (n=10,488) to build the models; 2) a testing data set comprising 15% of the data (n=2,247) to choose the optimal model among candidate models within a given class of model based on the estimated prediction error; and 3) a validation data set comprising 15% of the data (n=2,248) used to evaluate the final model predictions. The prediction errors in the testing data set were used to determine the final set of explanatory variables to input into the model.

Building a CART model requires the user to define tuning parameters that control the tree size. We kept constant values for the minimum number of jobs to allow a split within a node (20), the minimum number of jobs within a leaf (7), and the maximum node depth (30), the default settings in the rpart package. For each metric, we examined complexity parameters ranging from 0.0001 (most complex model) to 0.1 (simplest model). We selected the model with the lowest relative cross-validated error using 10-fold cross-validation in the training data set [8, 23, 24] to prevent the tree from overfitting the training data at the expense of the fit of the testing and validation data.

Random forests models combine many trees, where each tree is built from a random sample of the training data. The data left out in the training data when building any particular tree is referred to as the out-of-bag sample. An average prediction error for a random forest model can be calculated from averaging the prediction error from each of the hundreds of trees' out-of-bag sample. We used 300 trees because the average prediction error stabilized in the out-of-bag data after 100 trees. We used the square root of the number of input variables as the number of variables to consider at each split when building the individual trees [24]. We set the complexity parameter in the random forest model to the same value used for the best identified CART model.

Model evaluation

We evaluated the predictions of the best identified CART and random forest against the expert assignments within the validation data set based on the overall agreement with the expert's assignments and the percent agreement for each exposure category.

We conducted two sensitivity analyses. First, we examined the agreement of the models' predictions compared to the expert's assignments for models restricted to only occupational history variables (including the two supplementary diesel exposure questions), rather than all potential variables, to examine the reliability of the models' predictions when fewer or non-specific occupational data is collected. Second, we examined the sensitivity of the prediction reliability of the CART model (with complexity parameter = 0.01) to the size of the training data set by systematically increasing the number of jobs used in the training set in 5% increments to determine if we could reliably predict exposure assignments if we had exposure decisions for only a subset of the data. For each training set size, the prediction error was calculated based on the remaining data. We resampled 100 training sets of each size to estimate the distribution of prediction errors.

RESULTS

Decision rules

We first present a simple CART model that classified jobs into binary probability categories that was user-constrained to a high value (0.01) for the complexity parameter to limit the growth of the tree (Figure 2). The tree had a prediction agreement of 93.2% in the validation

set. The most predictive variable was `worked near or smelled exhaust', which was constructed from the two diesel exhaust-related questions in the occupational histories.

We observed somewhat better agreement when we allowed the tree to grow larger. Variables identified in decision rules for these more complex CART models are listed in Supplemental Material, Table 1. Decision rules from the CART models are available by contacting the corresponding author.

Model performance

Binary probability—For binary exposure, both the CART and random forest models exhibited a high overall agreement with the expert's rating (92–94%) in the validation data set (Table 1). Higher agreement was observed in jobs assessed as negligible/low exposed by the expert (93–95%) and lower agreement was observed in jobs assessed as medium/high exposed (79–92%). The models restricted to the occupational history variables had about 10% lower agreement in the medium/high category than the full models, but the two models had similar agreement in the negligible/low category. The random forest and CART models had the same agreement in the negligible/low category, but the agreement for the medium/high category was 1–4% higher in the random forest models than the CART models.

Ordinal probability—For ordinal exposure probability, the CART and random forests models' predictions agreed with 85–89% of the expert's assignments (Table 1). The agreement was highest (97–98%) for jobs assessed as unexposed by the expert. For jobs with a high rating, the agreement dropped from 85% when all variables were used to 68–72% when restricted to occupational history variables. The agreement was 7–43% for jobs assessed by the expert as having a low or medium rating. CART models had higher agreement with the expert's assignments than random forest models for jobs assessed as having low (32% versus 23%) and medium ratings (21% vs. 14%) when all variables were used. Poorer agreement was observed in the categories with lower prevalence.

Intensity—For exposure intensity, the CART and random forests models predictions' agreed with 87–90% of the expert's assignments (Table 2). Both CART and random forest models predicted jobs with no exposure well (agreement 96–98%) and had moderate to moderately high agreement with the expert ratings for jobs with low intensity (64–71%), medium intensity (41–57%), and high intensity (60–65%).

Frequency—For exposure frequency, the CART and random forests models' predictions agreed with 83–87% of the expert's assignments (Table 2). The predictions for jobs assessed as having no frequency of exposure agreed well (97–98%) with the expert's assignments. The agreement was poor to moderate for jobs rated as low (26–52%) or medium (12–39%) frequency, and moderate for jobs rated as high frequency (57–65%). Agreement was consistently higher for the models fit using all variables compared to using only the occupational history variables, but no consistent pattern was observed for the random forest models compared to the CART models.

Pattern of disagreements—The cross-tabulations of the predicted estimates for probability, intensity, and frequency from the CART model compared to the expert's estimates are shown in Table 2 for the validation data set. Similar patterns were observed for these three metrics. When a disagreement occurred, the CART model tended to predict a lower exposure rating than the expert for the two middle categories. It was rare for the CART model to predict a medium or high exposure rating when the expert assigned an unexposed rating (e.g., probability metric: 23 jobs) or for the CART model to predict a low

or unexposed rating when the expert assigned a high rating (e.g., probability metric: 40 jobs). Similar patterns were observed for the random forest models (not shown).

Training set size—The CART model's prediction error generally decreased as the number of jobs used in the training data set increased, with a plateau occurring when at least 3,750 randomly chosen jobs (25% of the nearly 15,000 jobs) were used (Figure 3). The largest median validation error occurred when using 5% of the data for training. The variance in prediction error was generally largest at the extreme training sample sizes (5%, 95%), where there were few jobs to use to either train the model or to evaluate the model performance, respectively.

DISCUSSION

We applied statistical learning methods to explain and predict an expert's exposure estimates derived from subjects' responses to an occupational questionnaire in a case-control study of occupational diesel exhaust exposure and bladder cancer. We found that the models had excellent ability to reproduce the expert's assignments for a binary probability metric and for the unexposed category for three ordinal metrics. For the ordinal metrics the models had poor to moderately-high ability to reproduce the experts' assignments for the exposed categories. However, the models identified the groups of questionnaire response patterns where agreement was poor. Thus, we recommend a two-stage assignment process to apply the resulting decision rules to unassessed jobs: initial assignment of decision rules, followed by expert review of the jobs identified by the model as more difficult to correctly classify.

Our CART and random forest models had similar predictive ability to that of artificial neural network (ANN) models used by Black et al. [25] to predict a dichotomous exposure for benzene exposure. ANN models are also a statistical learning approach, but ANNs use internal weights that can not be easily reviewed for plausibility by outside experts [10, 11]. We used, instead, tree-based methods such as CART because tree-based models provide both a visual and easily understood set of rules underlying the expert's exposure decisions. The extracted rules do not necessarily represent the decision process used by the expert. Instead, the models' rules extract questionnaire response patterns that best predict an expert's exposure decision. Black and colleagues [25] suggest that 60% of assessor's time can be saved by application of ANN models to identify unexposed jobs. We also predict a substantial reduction in the exposure assessment burden from using CART models to assign exposure in subsequent studies. Exposure assessors can focus their efforts on evaluating jobs that the tree-based methods found more difficult to classify, such as when the probability assessment for a leaf straddles the assignment cut point (e.g., 20–80%).

While random forests generally outperform CART models in prediction [8], the CART and random forest approaches used here performed similarly in predicting the expert's assignments. Any slight reduction in performance of the CART models compared to the random forest models is a tradeoff for the CART models' greatly improved interpretability, i.e., having only one decision tree rather than hundreds. The CART models' predictive abilities across the exposed categories might be improved further if the ordinal nature of the exposure metrics is considered instead of the categorical treatment used in the functions called by the R package rattle.

Overall, 66 of the 498 occupational history variables and 40 of the 223 module variables were predictive in a CART model for at least one exposure metric (Supplemental Material, Table 1). Coding the occupational history questions was a time-consuming, but essential step to developing the input variables for the models and required an occupational health professional. Without this, our potential explanatory determinants from the occupational

histories would have been restricted to SOC, SIC, job start and stop years, and the two supplementary questions, whereas the extracted decision rules revealed that the coded occupational history variables were important determinants. Limiting the models to using only the occupational history variables had little effect on the ability to reproduce the expert's classification of jobs as unexposed, but generally decreased the ability to reproduce the expert's classification of jobs into exposed categories. In our study, the adequate predictive ability of the models based only on the occupational histories likely reflects the use of the two engine/exhaust-related questions into the occupational histories, because the constructed variable derived from these two study-specific questions generally appeared to be the most predictive variable in all models. These two questions represent, in part, the subjects' self-assessment of exposure. However, these questions related to all types of engines and exhausts and not solely diesel exhaust and thus the self-assessment was not a perfect predictor of exposure status. The classification trees revealed that the expert's review considered whether there was additional supporting information for diesel exhaust sources in the responses. The increased ability of the models when using all variables (the occupational history and modules) to replicate the expert's assignments provides support for using modules in population-based studies to capture important within-job differences, but using modules can require a substantial time burden on the study participant and a substantial study cost to the interview and exposure assessment. The extracted decision rules, however, have identified the most important diesel-related questions, which can be used to simplify subsequent questionnaires for similar populations. This may reduce participant burden without losing much in the ability to reliably assign exposures using expert judgment.

Our sensitivity analyses revealed only small differences in performance between models. For example, overall agreement improved only 1% for the binary probability metric when the complexity parameter moved from 0.01 to 0.0006, although the number of rules needed to explain the model increased from 11 to 55, indicating that even a simple model was able to predict the binary exposure status well. Similarly, when we varied the number of jobs used in the training data set, the prediction error plateaued for all metrics when at least 25% (3,750 jobs) of the jobs was used. This suggests that an expert may be able to assess a random subset of the jobs after which CART models can be developed to provide reliable exposure predictions for the remaining jobs. The required size may, however, vary based on the number of jobs, prevalence of exposure, and the predictive ability of the model. Additional sensitivity analyses could be used to evaluate the appropriate minima for the number of observations per node and leaf. Some important determinants may not be captured using CART models when few subjects answered a particular question or when the minima are set too high. Thus, the determinants extracted here reflect common, not rare, exposure scenarios. Our focus on specificity rather than sensitivity in capturing determinants is appropriate because high specificity generally minimizes the expected attenuation in exposure-response associations when exposure prevalence is low [26].

The predictions of CART and random forest models are likely only as valid as the expert's exposure assessment [27], although the models could reduce some error from inconsistently applied rules. Therefore, our measures of agreement do not provide reassurance that exposures are classified correctly and provide only insights into reliability of the estimates if a CART model was used to assign exposures instead of one-by-one expert assessment. However, because no gold standards exist, extracting these rules provides an important first step to opening the black box to provide transparent decision rules so that other exposure assessors can review and revise the rules and thereby improve the quality of the assessments. Review of the models can be used to recognize discrepancies within the expert's estimates and to determine whether the decision rules could be improved by identifying additional explanatory variables, whether the condition is so rare that it cannot be captured in the model, or whether the expert's estimates should be improved [28]. After these improvements

are made from the internal and external reviews, new models can be developed to improve upon previous models.

Extrapolation outside of the scope of the study should be done carefully and may require important modifications to secular and geographic trends in exposure. Diesel exhaust exposure may also represent a best-case scenario. Diesel exhaust exposure is relatively common compared with other exposure agents often evaluated within population-based case-control studies, and it may be easier for subjects to identify and recall due to familiarity of diesel exhaust in the general population. In addition, the job- and industry-specific modules used in this study were specifically developed to collect information on diesel exhaust exposure. Our study provides a first step in demonstrating that CART models are able to extract underlying patterns between questionnaire responses and an expert's ratings for an agent that had been the focus of the questionnaires and had a reasonable prevalence rate. Future evaluations are needed to examine the utility of these models to extract decision rules for other agents that have lower exposure prevalences and that were not the primary focus. In addition, future evaluations are needed to determine whether similar decision rules would be extracted from exposure estimates provided by multiple independent or panels of experts.

Statistical learning approaches, such as CART and random forest models, offer great promise for explaining and predicting expert-based exposure estimates. Our approach was specific to extracting decision rules from previously made exposure assessments and can only be used in settings that are similar to those of where the decision rules were derived. For evaluating new exposures or new settings, we encourage exposure assessors to develop deterministic rules based on the questionnaire responses and program transparent assessments [1, 6]. The statistical learning approaches used here are straightforward to apply using the free GUI (rattle) within R, making these approaches accessible. The resulting models had excellent specificity allowing expert assessment of unexposed jobs to be reproduced with a great level of fidelity. The sensitivity was generally moderate, but, nonetheless, could reduce review time, especially if the models' estimates of the probability of belonging to an exposure category were used to triage jobs for further expert review. We encourage other researchers to apply these types of models to expert-based exposure assessments to describe the underlying decision rules. This will provide important insights into the rationales for exposure decisions, identify where exposure decisions may be inconsistent, and identify the most important information used by the expert to make an exposure decision. Building this body of knowledge will allow us to refine questionnaires to reduce subject burden and more rapidly provide exposure estimates in subsequent studies to test the reproducibility of findings across populations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding The research was funded by the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Division of Cancer Epidemiology and Genetics.

REFERENCES

1. Fritschi L, Friesen MC, Glass D, et al. OccIDEAS: Retrospective occupational exposure assessment in community-based studies made easier. Journal of Environmental and Public Health. 2009:2009.
2. Gerin M, Siemiatycki J, Kemper H, et al. Obtaining occupational exposure histories in epidemiologic case-control studies. J Occup Med. 1985; 27(6):420–6. [PubMed: 4020500]

3. Stewart PA, Stewart WF, Siemiatycki J, et al. Questionnaires for collecting detailed occupational information for community-based case control studies. Am Ind Hyg Assoc J. 1998; 59(1):39–44. [PubMed: 9438334]
4. Kauppinen T. Exposure assessment--a challenge for occupational epidemiology. Scand J Work Environ Health. 1996; 22(6):401–3. [PubMed: 9000306]
5. Kromhout H. Commentary. Occupational and Environmental Medicine. 2002; 59(9):594.
6. Pronk, A.; Stewart, PA.; Coble, JB., et al. Comparison of two expert-based assessments of diesel exhaust exposure in a case-control study: decision algorithms versus expert review of individual jobs. Submitted
7. Berk, R. Statistical learning from a regression perspective. Springer; New York: 2008.
8. Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning. Springer-Verlag; New York: 2001.
9. Colt J, Karagas M, Schwenn M, et al. Occupation and bladder cancer in a population-based case-control study in Northern New England. Occupational and Environmental Medicine. 2011; 68:239–49. [PubMed: 20864470]
10. Kennedy SM, Le Moual N, Choudat D, et al. Development of an asthma specific job exposure matrix and its application in the epidemiological study of genetics and environment in asthma (EGEA). Occup Environ Med. 2000; 57(9):635–41. [PubMed: 10935945]
11. Flouris A, Duffy J. Applications of artificial intelligence systems in the analysis of epidemiological data. Eur J Epidemiol. 2006; 21(3):167–70. [PubMed: 16547830]
12. Meyfroidt G, Guiza F, Ramon J, et al. Machine learning techniques to examine large patient databases. Best Pract Res Clin Anaesthesiol. 2009; 23(1):127–43. [PubMed: 19449621]
13. Breiman, L.; Friedman, J.; Olshen, R., et al. Classification and regression trees: Wadsworth & Brooks/Cole Advanced Books & Software. 1984.
14. Breiman L. Random forests. Machine Learning. 2001; 45(1):5–32.
15. Dietterich T. Ensemble methods in machine learning. Lecture Notes in Computer Science. 2000; 1857:1–15.
16. Office of Management and Budget. Standard industrial classification manual. Executive Office of the President; Washington, D.C: 1987.
17. U.S. Department of Commerce. Standard occupational classification manual. Office of Federal Statistical Policy and Standards; Washington, D.C.: 1980.
18. Williams GJ. Rattle: a data mining GUI for R. The R Journal. 2009; 1/2:45–55.
19. Williams, GJ. rattle: a graphical user interface for data mining in R. R package version 2.6.6 ed. 2009.
20. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2006.
21. Therneau, T.; Atkinson, B. rpart: recursive partitioning. R package version 3.1-46 ed. 2010.
22. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002; 2(3)
23. Torgo, L. Data mining with R: learning with case studies. Chapman & Hall/CRC; Boca Raton, FL: 2011.
24. Williams, GJ. Data mining with Rattle and R. Springer; New York: 2011.
25. Black J, Benke G, Smith K, et al. Artificial neural networks and job-specific modules to assess occupational exposure. Ann Occup Hyg. 2004; 48(7):595–600. [PubMed: 15381511]
26. Dosemeci M, Stewart P. Recommendations for reducing the effects of exposure misclassification on relative risk estimates. Occupational Hygiene. 1996; 3:169–76.
27. Burstyn I. The ghost of methods past: exposure assessment versus job-exposure matrix studies. Occup Environ Med. 2011; 68(1):2–3. [PubMed: 21075766]
28. Friesen, MC.; Pronk, A.; Wheeler, DC., et al. Comparison of algorithm-based estimates of occupational diesel exhaust exposure to those of multiple independent raters in a population-based case-control study. submitted

What this paper adds

- Expert-based exposure assessment of occupational risk factors in population-based case-control studies is challenging, time-consuming, and is criticized for its lack of transparency. Evaluating exposures in these studies often requires exposure assessors to review each study participants' reported occupational information job-by-job to derive exposure estimates.
- The structured format of occupational history and job-specific modules in questionnaires, however, makes it possible to identify underlying expert exposure decision rules.
- The present study is the first to use the statistical learning techniques of classification and regression trees and random forests to identify the underlying decision rules of an exposure assessor.
- The good agreement between the model predictions and one-by-one expert evaluations provides support for extracting transparent, identifiable decision rules from previously made expert assessments. This approach can be used to focus expert review efforts on questionnaire response patterns that had poorer prediction replication.

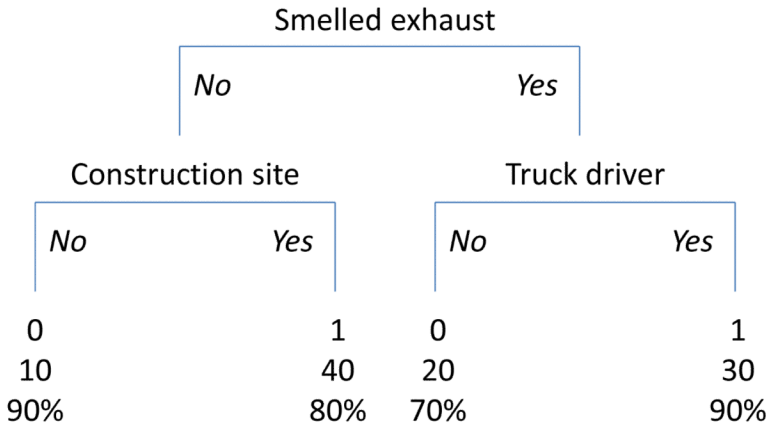

Figure 1.
Illustrative decision tree for the classification of 100 jobs by diesel exhaust exposure. The terminal nodes at the bottom of the tree are leaves with labels for exposure classification (0 = unexposed, 1 = exposed), number of jobs in leaf, and percent agreement of tree-based classifications with exposure status assigned by an expert.

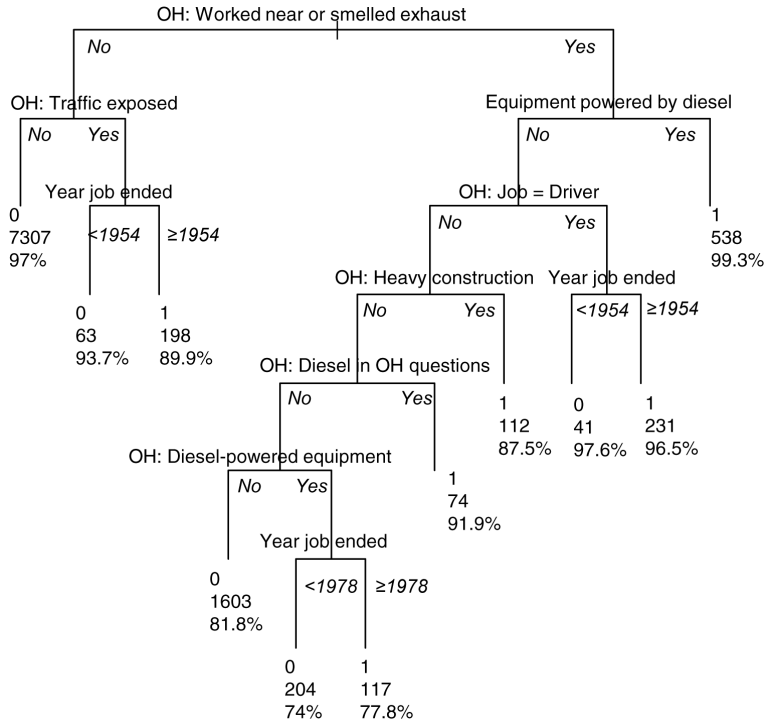

Figure 2.

CART decision tree classifying jobs into exposed (0) and unexposed (1) categories. The labels in each leaf, in order, are the predicted exposure category, the number of jobs in the leaf, and the percent of predictions in the leaf that agree with the expert estimate. Variables from the occupational history are designated OH; variables from the modules are designated M.

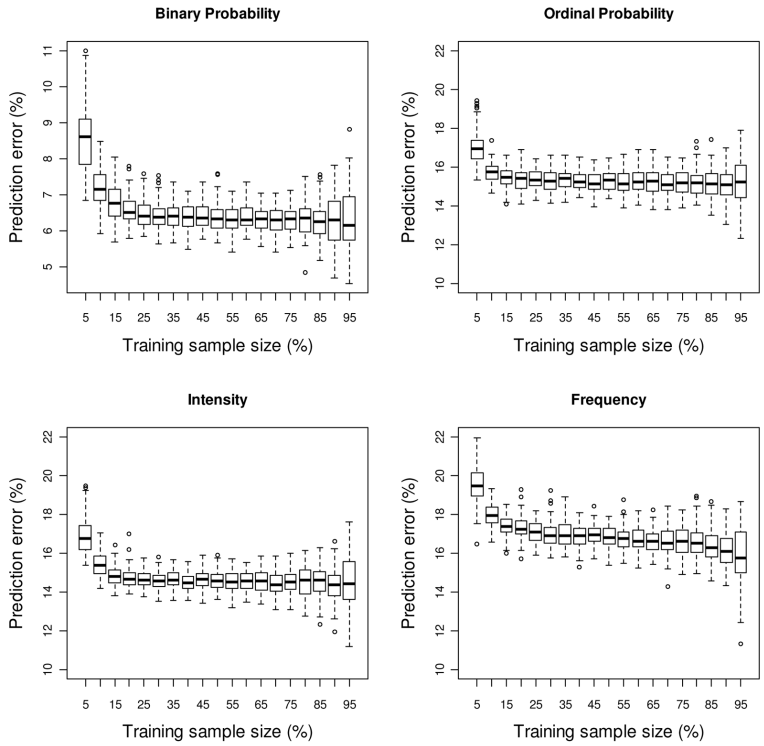

Figure 3.

CART prediction errors in the validation data set as the size of the training set varies for four exposure metrics: binary exposure probability, ordinal probability, intensity, and frequency of exposure. Each boxplot is based on 100 randomly selected training sets to estimate the model using all variables (complexity parameter = 0.01), with the prediction error estimated on the validation set.

Table 1

Proportion of exposure predictions from CART and random forest models that agreed with the expert exposure estimate in the validation data set (n=2,248).

Expert Exposure Metric	Number of Jobs	Agreement (%)			
		CART, OH variables	CART, OH & module variables	Random Forest, OH variables	Random Forest, OH & module variables
Probability, Binary					
Negligible/Low	1887	93	95	93	95
Medium/High	358	79	89	80	92
Overall	2245	*92*	*94*	*92*	*94*
Probability, Ordinal					
Negligible	1705	97	97	98	98
Low	182	32	43	23	32
Medium	73	7	21	8	14
High	285	68	85	72	85
Overall	*2245*	*85*	*89*	*85*	*89*
Intensity					
None	1708[a]	98	97	96	97
Low	394	64	68	67	71
Medium	86	48	56	41	57
High	57	61	65	61	60
Overall	*2245*	*87*	*89*	*88*	*90*
Frequency					
None	1757[a]	97	97	98	98
Low	209	32	52	26	34
Medium	141	12	39	13	35
High	141	57	63	59	65
Overall	*2248*[b]	*83*	*87*	*84*	*86*

CART, classification and regression tree; OH, occupational history.

[a] The number of unexposed jobs based on intensity and frequency exceeds the number of unexposed jobs based on the probability metric. This occurred because the estimated level of exposure intensity and frequency for some jobs did not exceed the minimum threshold (<0.25 $\mu g\ m^{-3}$ REC and < 0.25 hours per week, respectively), even though a diesel exposure source was identified.

[b] There were 3 additional observations for the frequency metric; these observations were excluded from the probability and intensity analyses because for those metrics an `unknown' had been assigned.

Table 2

Cross-tabulation of the CART model-predicted assignments versus expert assignments and proportion of predicted estimates that agreed with expert assignments in the validation data set (n=2,248).

Expert	CART Model[a]				
Exposure Metric	Predicted Estimate, Number of Jobs				Agreement (%)
	None	Low	Medium	High	
Probability					
Negligible	1660	22	4	19	97
Low	87	78	0	17	43
Medium	32	11	15	15	21
High	34	6	3	242	85
Intensity					
None	1651	52	2	3	97
Low	115	269	8	2	68
Medium	13	21	48	4	56
High	15	6	0	36	65
Frequency					
None	1700	32	15	10	97
Low	85	109	10	5	52
Medium	53	12	55	21	39
High	25	19	8	89	63

[a] All variables offered in models.