Ann. Occup. Hyg., 2015, Vol. 59, No. 3, 324–335
doi:10.1093/annhyg/meu098
Advance Access publication 27 November 2014

Comparison of Ordinal and Nominal Classification Trees to Predict Ordinal Expert-Based Occupational Exposure Estimates in a Case–Control Study

David C. Wheeler[1]*, Kellie J. Archer[1], Igor Burstyn[2], Kai Yu[3], Patricia A. Stewart[4], Joanne S. Colt[5], Dalsu Baris[5], Margaret R. Karagas[6], Molly Schwenn[7], Alison Johnson[8], Karla Armenti[9], Debra T. Silverman[5] and Melissa C. Friesen[5]

1.Department of Biostatistics, School of Medicine, Virginia Commonwealth University, 830 East Main Street, Richmond, VA 23298, USA
2.Drexel University, School of Public Health, Nesbitt Hall, 3215 Market Street, Philadelphia, PA 19104, USA
3.Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 9609 Medical Center Drive, MSC 9776, Bethesda, MD 20892, USA
4.Stewart Exposure Assessments, LLC, 6045 27th Street North, Arlington, VA 22207, USA
5.Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 9609 Medical Center Drive, MSC 9776, Bethesda, MD 20892, USA
6.Geisel School of Medicine at Dartmouth, 1 Medical Center Drive, 7927 Rubin Building, Lebanon NH 03756, USA
7.Maine Cancer Registry, 286 Water Street, 4th Floor, 11 State House Station, Augusta, Maine 04333-0011, USA
8.Vermont Cancer Registry, Vermont Department of Health, P.O. Box 70, Burlington, VT 05402-0070, USA
9.New Hampshire Department of Health and Human Services, 29 Hazen Drive, Concord, NH 03301, USA
*Author to whom correspondence should be addressed. Tel: +1-804-828-9827; fax: +1-804-828-8900; e-mail: dcwheels@gmail.com

Submitted 11 July 2014; revised 16 October 2014; revised version accepted 20 October 2014.

ABSTRACT

Objectives: To evaluate occupational exposures in case–control studies, exposure assessors typically review each job individually to assign exposure estimates. This process lacks transparency and does not provide a mechanism for recreating the decision rules in other studies. In our previous work, nominal (unordered categorical) classification trees (CTs) generally successfully predicted expert-assessed ordinal exposure estimates (i.e. none, low, medium, high) derived from occupational questionnaire responses, but room for improvement remained. Our objective was to determine if using recently developed ordinal CTs would improve the performance of nominal trees in predicting ordinal occupational diesel exhaust exposure estimates in a case–control study.

Methods: We used one nominal and four ordinal CT methods to predict expert-assessed probability, intensity, and frequency estimates of occupational diesel exhaust exposure (each categorized as none, low, medium, or high) derived from questionnaire responses for the 14983 jobs in the New England Bladder Cancer Study. To replicate the common use of a single tree, we applied each method to a single sample of 70% of the jobs, using 15% to test and 15% to validate each method. To characterize variability in performance, we conducted a resampling analysis that repeated the sample draws 100 times. We evaluated agreement between the tree predictions and expert estimates using Somers' *d*, which measures differences in terms of ordinal association between predicted and observed scores and can be interpreted similarly to a correlation coefficient.

Results: From the resampling analysis, compared with the nominal tree, an ordinal CT method that used a quadratic misclassification function and controlled tree size based on total misclassification cost had a slightly better predictive performance that was statistically significant for the frequency metric (Somers' *d*: nominal tree = 0.61; ordinal tree = 0.63) and similar performance for the probability (nominal = 0.65; ordinal = 0.66) and intensity (nominal = 0.65; ordinal = 0.65) metrics. The best ordinal CT predicted fewer cases of large disagreement with the expert assessments (i.e. no exposure predicted for a job with high exposure and vice versa) compared with the nominal tree across all of the exposure metrics. For example, the percent of jobs with expert-assigned high intensity of exposure that the model predicted as no exposure was 29% for the nominal tree and 22% for the best ordinal tree.

Conclusions: The overall agreements were similar across CT models; however, the use of ordinal models reduced the magnitude of the discrepancy when disagreements occurred. As the best performing model can vary by situation, researchers should consider evaluating multiple CT methods to maximize the predictive performance within their data.

KEYWORDS: classification; diesel exhaust; occupational exposure; ordinal data; statistical learning

INTRODUCTION

To derive occupational exposure estimates in case–control studies of cancer and other chronic diseases, exposure assessors often review the occupational information for each job reported in questionnaires by study participants. Though such assessments have underlying decision rules, they usually lack transparency and are time-consuming to conduct (Wheeler *et al.*, 2013). Fortunately, there is growing evidence that using rule-based approaches results in similar exposure estimates to a one-by-one review (Behrens *et al.*, 2012; Pronk *et al.*, 2012; Friesen *et al.*, 2013; Wheeler *et al.*, 2013; Carey *et al.*, 2014; Peters *et al.*, 2014) and may reasonably reflect the average rating of multiple raters (Friesen *et al.*, 2013). Advantages of rule-based approaches include transparency and automation of the decisions, which allow for sensitivity analyses to be conducted and for feedback from other experts. Once decision rules have been developed, rule-based approaches can reduce the exposure assessment time for that agent in future studies.

We previously used classification trees (CTs) to efficiently identify the underlying assessment decision rules that explained patterns between occupational questionnaire responses and expert-assessed ordinal estimates of the probability, intensity, and frequency of occupational exposure to diesel exhaust in a bladder cancer case–control study (Wheeler *et al.*, 2013). Each job in the study was assigned an estimate for each metric, categorized as none, low, medium, or high. The CTs used in our previous study treated the exposures as nominal categorical variables (unordered) because of the limited availability of methods that treated the outcome as ordinal (ordered). We found that, despite using nominal classification methods, the agreement between the CT predictions and the expert assessments ranged from good to excellent for unexposed (86–90%) and highly exposed categories (57–85%). However, agreement was considerably less consistent and, at times, poor for low or medium exposed categories (7–71%). Though the performance was good overall, there was room for improvement in predicting the low and medium categories of exposure, particularly for frequency of exposure.

An important area for potential improvement would be to take into account the ordinal nature of the exposure metrics. For example, a previous simulation study showed that ensembles of CTs designed for nominal data had higher prediction error and lower agreement with an ordinal outcome according to a gamma statistic (an ordinal-based evaluation metric) than several CT ensembles designed for ordinal outcomes (Archer and Mas, 2009). Moreover, in the ordinal outcome setting, misclassifying observations to adjacent categories is a less egregious error than misclassifying observations to more distant categories. Nominal CTs do not take advantage of the additional information present in the ordinal response. Hence, our previous analysis using nominal CTs treated a one-category difference in assignment the same degree of disagreement as a two- or three-category difference in assignment. Because ordinal CTs account for the ordered nature of the categories, ordinal CTs are less likely to misclassify an observation into a distant category. The previous analyses were facilitated by readily available software for CTs for nominal categorical outcomes (Williams 2009; Therneau, Atkinson, Ripley, 2014) in the R computing environment (R Core Team, 2014). Recently, software for CTs designed for ordinal outcomes also became freely available (Archer 2010; Galimberti *et al.*, 2012) and is now practical to implement.

Our primary objective was to determine whether using CT methods designed specifically for ordinal outcome measures would improve the performance of nominal CTs in predicting ordinal diesel exhaust exposures in a case–control study. Our secondary objective was to provide an overview of the similarities and differences in CT methods for an exposure science audience. Using the occupational diesel exhaust example (Wheeler *et al.*, 2013), we compared the predictive performance between a previously used nominal CT method and four ordinal CT methods to predict an expert's diesel exhaust exposure estimates derived from questionnaires. First, we evaluated the performance between methods based on a single tree to replicate the common use of a single tree to predict exposure class (hereafter, single tree analysis). Second, we evaluated the variability in performance between methods using a resampling analysis where the tree-building process was repeated 100 times (hereafter, resampling analysis).

METHODS

Study population

Our study population was the previously used New England Bladder Cancer Study, where an expert assessed the probability, intensity, and frequency of occupational diesel exhaust exposure (Pronk *et al.*, 2012; Wheeler *et al.*, 2013). The study was composed of 1170 bladder cancer cases and 1413 controls in Maine, New Hampshire, and Vermont enumerated between 2001 and 2004. All respondents gave written informed consent to participate in this study. The study protocol was approved by the National Cancer Institute Special Studies Institutional Review Board, as well as the human subjects review boards of each participating institution. There were 14 983 total jobs reported by study participants using a lifetime occupational history questionnaire. A subset of jobs (64%) also received job- or industry-specific questionnaires that asked more detailed task and exposure information on a number of agents, including diesel exhaust.

The extraction of variables from the occupational questionnaires that were used in the previous and current CT model evaluation was described in detail in Wheeler *et al.* (2013). Briefly, from the occupational histories, we extracted or derived 498 variables related to possible diesel exhaust-exposed scenarios based on responses to questions on occupation, industry, main tasks or activities, and tools, equipment, materials, and chemicals used. These variables included 'job had traffic exposure', 'job used diesel equipment', 'industry likely or plausibly used diesel equipment', 'smelled or worked near diesel or other engines', and previously assigned standardized occupation and industry codes (Colt *et al.*, 2011). From the job- and industry-specific questionnaires, we extracted or derived an additional 223 diesel-related variables, including 'used diesel-powered equipment', 'smelled exhaust', 'worked near idling diesel equipment', 'repaired diesel equipment', 'equipment type', and 'job function'. These 721 variables were considered possible predictive variables.

The exposure estimates were previously obtained from a one-by-one review of each job by an industrial hygienist (P.A.S.) to assign the probability, intensity, and frequency of diesel exhaust exposure (Pronk *et al.*, 2012). Probability was assessed as the estimated proportion of workers likely exposed to diesel exhaust based on all the information in the occupational history and job- and industry-specific questionnaires, including tasks, job, industry, and decade, with estimated cut points of <5% (none/negligible, category 0), 5–49% (low, 1), 50–79% (medium, 2), and ≥80% of workers (high, 3). Approximately 75% of the jobs were assessed as having negligible probability of exposure. For all jobs assigned a probability ≥5%, intensity and frequency were estimated. Intensity was assessed on a continuous scale as the estimated average level of respirable elemental carbon (REC, μg m^{-3}) in the worker's breathing zone during tasks where diesel exhaust exposure occurred and, for the CT models, 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) μg m^{-3} REC. Frequency was assessed on a continuous scale as the estimated average number of hours per week exposed to diesel exhaust and, for the CT models, categorized with cut points of <0.25 (none/negligible, category 0), 0.25 to <8 (low, 1), 8 to <20 (medium, 2), and ≥20 (high, 3) hours per week. These cut points were set to the same categories defined for the original nominal CT models (Wheeler *et al.*, 2013). These categories are described below as the outcome (variable) class or score.

Statistical methods

CTs are built by recursively partitioning the data set using splitting rules, which are logical rules defined according to the values of selected explanatory variables. When deriving a CT, all observations in the training data start together in the root node, where nodes are denoted as t. Then, for each of the p predictor variables, the optimal binary split is determined. In node impurity-based CTs (impurity functions are defined below), optimality is defined as that split resulting in the largest decrease in node impurity, which is a measure of heterogeneity in the node with respect to the outcome variable class. Splits resulting in increasingly more homogeneous nodes with respect to class are preferred. Hence, the splitting rules are dependent on the node impurity function in this type of CT. Among all best splits for the p predictor variables, the very best split among the variables is selected for partitioning the observations to the left and right descendant nodes. This splitting process is repeated for all descendant nodes until the terminal nodes are either homogeneous with respect to the outcome class or there are too few observations for further partitioning according to a stopping rule. While splitting rules are used to build the tree, pruning and stopping rules, which vary based on the CT model chosen (described below) and user-specified parameters (e.g. minimum number of observations), control the growth or size of the tree. Pruning involves trimming some of the terminal nodes from a tree to prevent overfitting in a training data set, resulting in a smaller tree that may have better predictive performance in a validation data set. Both the splitting rules and pruning or stopping rules can impact the predictive performance of CT methods.

Here, we used the nominal CT method previously used (Wheeler *et al.*, 2013) and four ordinal CT methods to predict the expert-assessed exposure metrics for each job (dependent variables) using the extracted and coded variables from the occupational questionnaire responses (explanatory variables). The splitting and pruning or stopping rules for the five CT methods used here (nominal: rpart; ordinal: rpartScore MC, rpartScore MR, ctree Bonferroni, ctree Univariate) are described below. We used the following R packages: rpart (Therneau, Atkinson, Ripley, 2014) for CTs with a nominal Gini impurity function, rpartScore (Galimberti *et al.*, 2012) for CTs with a generalized Gini impurity function, and ctree (Hothorn *et al.*, 2006) for conditional inference trees. No particular ordinal method was expected *a priori* to provide substantially better performance because there has been no comparison of these ordinal models to each other in the published exposure assessment literature. Based on a previous analysis in the literature (Archer and Mas, 2009), we anticipated that at least the CT with a generalized Gini impurity function (rpartScore MC and/or rpartScore MR) would classify the ordinal diesel exhaust exposure metrics better than the CT with the nominal Gini impurity function (rpart).

rpart and rpartScore CT methods

The rpart and the two rpartScore methods use recursive partitioning to build a tree using a Gini impurity function to determine each split in the tree. The optimal binary splitting rule for a given node in a tree is the one that results in the largest decrease in node impurity, as measured by a generalized impurity function (I) (Breiman *et al.*, 1984; Galimberti *et al.*, 2012) at a tree node t, defined as

$$I(t)=\sum_{j=1}^{L}\sum_{k=1}^{L}C(w_j\,|\,w_k)p(w_j\,|\,t)p(w_k\,|\,t), \qquad (1)$$

where $C(w_j\,|\,w_k)$ is the misclassification cost of assigning an observation to category j of variable w when it belongs to category k, and $p(w_j\,|\,t)$ is the proportion of observations in node t that belong to category j of the L number of categories of the outcome variable. The nominal Gini impurity function used by rpart sets the misclassification cost $C(w_j | w_k) = 1$ when $j \neq k$ and sets $C(w_j | w_k) = 0$ when $j = k$ (correct assignment).

For both rpartScore methods, the misclassification cost is a measure of dissimilarity between the actual and assigned categories. Here, we used a quadratic misclassification function $C(w_j | w_k) = (s_j - s_k)^2$, where s_j is the classification score for category j of the outcome variable and s_k is the score for the observed category k using the ordinal coding scheme. In our case, the scores = {0,1,2,3} and the misclassification cost is the squared difference in category levels. For example, the misclassification cost for the method assigning a job assessed by the expert as having a low (1) exposure to a high (3) exposure category would be $(3 - 1)^2 = 4$; the misclassification cost for assigning the same job to no exposure would be $(0 - 1)^2 = 1$. As a result, the rpartScore CTs place more emphasis in the impurity function on observations that are incorrectly classified far from the true class.

The two rpartScore methods differ in the pruning function controlling the size of the tree. A tree that is too large may fit the training data very well but predict poorly in a testing data set. The pruning approach used here is the cost-complexity measure in equation (2) (Breiman *et al.*, 1984; Galimberti *et al.*, 2012), which combines a measure of predictive performance and a measure of the complexity of the tree, usually the number of leaves or terminal nodes. Specifically, the cost-complexity measure is

$$R_\alpha(T) = R(T) + \alpha \times card(T), \qquad (2)$$

where $R(T)$ is the predictive performance, $card(T)$ is the size of the tree T measured by the number of terminal nodes, and α is a tuning parameter that controls the trade-off between predictive performance and model complexity. The rpartScore MC sets the predictive performance measure to the total misclassification cost, whereas rpartScore MR sets it to the total number of misclassified observations. The latter is a sum of all the observations that are classified incorrectly, whereas the former is a sum over all observations of the absolute difference between the observed score and the predicted score.

To build and prune the rpart and rpartScore CTs, we selected the best model for each type of tree using the 1-SE rule (Breiman *et al.*, 1984) where the SE was estimated using 10-fold cross-validation in the testing set and a complexity parameter (α in equation (2)) ≥0.001, a small minimum bound to allow for pruning. The 1-SE rule selects the tree that has the maximum predictive error that is within 1 SE of the minimum predictive error.

ctree CT methods

The ctree methods are conditional inference trees, which estimate a regression relationship by binary recursive partitioning in a conditional inference framework (Hothorn *et al.*, 2006). The conditional inference tree approach begins by testing the global null hypothesis of independence between the outcome variable and any of the input variables. If the hypothesis is not rejected, the algorithm selects the input variable with the strongest association with the response. The association is measured by a *P*-value corresponding to a test for the partial null hypothesis of a single input variable and the outcome variable. A binary split on the selected predictor is then made. The process of testing for a significant association and then conducting a binary split on the selected significant predictor continues until a stopping criterion is satisfied. The condition for stopping is based on univariate *P*-values for the ctree Univariate method and on multiplicity-adjusted *P*-values for the ctree Bonferroni method. A split is made as long as the minimum univariate or Bonferroni-adjusted *P*-value is below a nominal level, such as 0.05. Due to the nature of the stopping criteria, no pruning is necessary using conditional inference trees. Ordinal outcome variables are accommodated through scores that reflect the distance between categories of the outcome.

Evaluating performance of CT methods

To evaluate the performance of the five CT methods in predicting diesel exhaust exposure score, we conducted a single tree analysis and a resampling analysis using each of the methods. The 498 occupational history variables and 223 job- and industry-specific variables were the predictors considered to build the CTs. For the single tree analysis, we used a single 70% sample (n = 10 488 jobs) to build the model, a 15% sample (n = 2247) to select a model among a set of candidate models, and a 15% sample (n = 2248) to validate the selected model. Candidate models differed in the set of input variables (e.g. including or excluding standardized occupation and industry codes) considered to

Table 1. Single tree analysis: Somers' *d* values based on the validation set comparing the predicted estimates from five CT models to the expert-assigned estimates for three diesel exhaust exposure metrics

Model[a]	Somers' *d*		
	Probability	Intensity	Frequency
rpart	0.67	0.66	0.60
rpartScore MC	0.70	0.67	0.60
rpartScore MR	0.64	0.66	0.56
ctree Bonferroni	0.55	0.60	0.58
ctree Univariate	0.56	0.58	0.61

[a] See text for descriptions of the differences in splitting and stopping/pruning rules used in each model.

build the CT. For the resampling analysis, we generated 100 training samples (70% of data) and evaluated model performance in 100 validation sets (30% of data). The input variable set for the best model from the single tree analysis was used to build the CTs in the resampling analysis. In each of the 100 training samples, sampling of the jobs was stratified proportional to the number of jobs assigned by the industrial hygienist to each category of each exposure metric.

In both the single tree and resampling analyses, measures of the predictive performance of the models were evaluated on the validation set. We evaluated the agreement of the model predictions and the expert exposure assignments in the validation sets using Somers' *d* (Agresti, 2002), which measures differences in terms of ordinal association between predicted and observed scores (Galimberti *et al.*, 2012), and can be interpreted similarly to other correlation measures with a scale ranging from −1 to 1. It is an asymmetric measure of association, computed as the difference between the proportions of pairs of observations that are concordant and those that are discordant in the observed and predicted scores (among the observations that are not tied in the observed scores). We tested the global hypothesis of no difference in the agreement among models using Friedman's non-parametric rank test for repeated measurements in a randomized complete block design, treating each of the 100 training sets as a block (Hollander and Wolfe, 1999; Galimberti *et al.*, 2012). The test statistic and asymptotic *P*-value were calculated using the *coin* R package (Hothorn *et al.*, 2008). We tested for differences in model agreement for pairs of models using the Wilcoxon signed rank test, where the primary interest was on differences between the nominal tree and ordinal trees.

RESULTS

In the single tree analysis, the best agreement for probability of exposure and intensity of exposure was observed using rpartScore MC, with rpart only slightly lower (Table 1). The best agreement for frequency of exposure was with ctree Univariate; however, rpart and rpartScore MC provided nearly identical predictive performance. For the rpart and rpartScore methods, CTs for the probability metric had nearly the same or slightly better performance than CTs predicting intensity, with the lowest performance observed for CTs predicting frequency. For the ctree methods, the order from highest to lowest performance by metric was intensity, frequency, and probability for the Bonferroni stopping condition and frequency, intensity, and probability for univariate stopping.

To examine where certain methods differed in prediction agreement with the expert assessments, we calculated the absolute value of the difference in scores for the expert-assigned exposures and the CT predicted exposures. The cross-tabulations for the absolute score differences for the two best overall methods according to the single tree analysis (rpart and rpartScore MC) methods are listed in Table 2 for the three exposure metrics. The (0, 0) cell in the table for each metric lists the number of jobs where both rpart and rpartScore MC predictions agreed with the expert-assigned exposure. The (3, 0) cell

Table 2. Single tree analysis: cross-tabulation of the absolute differences in ordinal scores for expert-assigned exposures and predicted exposures from the nominal rpart and ordinal rpartScore MC CTs in the validation data set (n = 2248 jobs).

rpart	rpartScore MC			
	Absolute score difference between predicted and assigned category			
	0	1	2	3
Probability difference				
0	1930	48	6	1
1	26	116	6	0
2	4	11	49	1
3	6	5	5	34
Intensity difference				
0	1948	48	0	0
1	24	116	1	0
2	5	10	21	0
3	3	0	5	15
Frequency difference				
0	1882	40	8	5
1	12	161	5	0
2	9	8	71	1
3	6	11	3	26

lists the number of jobs where the rpart predicted score differed by 3 from the expert assignment, but the rpartScore MC predicted score matched the expert-based score. This is the situation where rpart predicted either no exposure for a high exposure job or a high exposure for a no exposure job, i.e. the largest possible disagreement. The table shows that rpartScore MC had fewer large disagreements in exposure estimates (score difference of 2 or more) than did rpart. The number of large disagreements for rpart compared with rpartScore MC was 15 (6 + 4 + 5) to 7 (1 + 6 + 0) for probability, 8 to 0 for intensity, and 26 to 13 for frequency. The tendency for rpartScore MC to have fewer large disagreements than rpart was also evident in tables of predicted versus expert-assigned exposure scores across the exposure metrics, even though the proportion in perfect agreement was usually slightly higher with rpart than rpartScore MC (see Supplementary Table 1, available at *Annals of Occupational Hygiene* online).

To compare the composition of each tree, we determined the number of variables in each of the five tree models for each exposure metric and the number of variables that were common to each of the four ordinal trees and the nominal tree. There was considerable overlap between the variables selected by the ordinal trees and the nominal tree, but also some differences (Table 3). For example, rpartScore MC had 77% (109/141) of its variables in common with rpart for probability, 72% for intensity, and 79% for frequency of exposure. The rpartScore MC trees had fewer variables than the rpart trees for each exposure metric. The ctree models were substantially smaller than the other trees. There was also some overlap in the 10 most important variables for each tree (see Supplementary Tables 2 and 3, available at *Annals of Occupational Hygiene* online). For example, using equipment powered by diesel was either the most important or second most important variable in all trees for probability of exposure. Additional details on the variables for the nominal tree are available in Wheeler *et al.* (2013).

In the resampling analysis, similar patterns to the single tree analysis were observed in the mean Somers' d values comparing model and expert estimates (Table 4). For the rpart and rpartScore methods, probability and intensity had approximately similar performance measures, with rpartScore MC having better performance for frequency. The ctree methods were slightly lower (0.01–0.05 units) than the rpart and rpartScore methods, with the lowest performance observed for intensity. The two ctree methods were nearly indistinguishable in performance from each other, with the Univariate method having slightly better performance for the frequency measure than the Bonferroni method.

The above patterns are more clearly shown in the distributions of the Somers' d metrics (Fig. 1). For probability and intensity, the rpart and rpartScore distributions were similar to each other and were higher than both ctree models. For exposure frequency, the median agreement from rpartScore MC was the

Table 3. Number of variables selected in each tree model and the number of variables common to each model and the nominal CT (rpart) in the single tree analysis.

Model	Probability		Intensity		Frequency	
	Number variables	Common with rpart	Number variables	Common with rpart	Number variables	Common with rpart
rpart	166	166	140	140	161	161
rpartScore MC	141	109	130	93	122	96
rpartScore MR	154	120	162	102	154	108
ctree Bonferroni	19	18	27	24	21	21
ctree Univariate	26	20	37	29	44	38

Table 4. Resampling analysis (100 trees): mean Somers' *d* values based on the validation set comparing the predicted estimates from five CT models to the expert-assigned estimates for three diesel exhaust exposure metrics.

Model[a]	Somers' *d* (variance)		
	Probability	Intensity	Frequency
rpart	0.65 (0.0004)	0.65 (0.0004)	0.61 (0.0004)
rpartScore MC	0.66 (0.0004)	0.65 (0.0004)	0.63 (0.0004)
rpartScore MR	0.65 (0.0006)	0.65 (0.0004)	0.60 (0.0007)
ctree Bonferroni	0.62 (0.0003)	0.60 (0.0005)	0.61 (0.0007)
ctree Univariate	0.62 (0.0004)	0.60 (0.0004)	0.62 (0.0004)

[a] See text for descriptions of the differences in splitting and stopping/pruning rules used in each model.

highest, and its distribution of agreement was generally higher those from the other models. For frequency, the interquartile range and overall range from rpartScore MC were considerably narrower than that from rpartScore MR and rpart. Generally, the two ctree methods had similar median and interquartile agreement measures.

We also examined the relationship between the Somers' *d* measures across the 100 models built in the resampling analysis for each combination of method pairs using scatter plots (see Supplementary Figures 1–3, available at *Annals of Occupational Hygiene* online). For frequency, the ctree methods were more similar to each other than to the rpart and rpartScore methods. Similarly, the two rpartScore methods were more similar to each other than the rpart and ctree models. Similar patterns were also observed for probability and intensity, although the two ctree methods were most correlated for the intensity metric.

Friedman's test rejected symmetry in agreement across the five models for each of the three exposure metrics ($P < 0.0001$). For exposure probability, the median agreement for rpart was significantly different from both ctree methods and rpartScore MR but was not significantly different from the rpartScore MC model according to the Wilcoxon tests. The median agreement for rpartScore MC was significantly different from that of the rpartScore MR and the ctree methods. For exposure intensity, the ctree models were significantly different in agreement from the three other models, but the rpart and rpartScore methods were not significantly different in agreement. For exposure frequency, there were significant differences

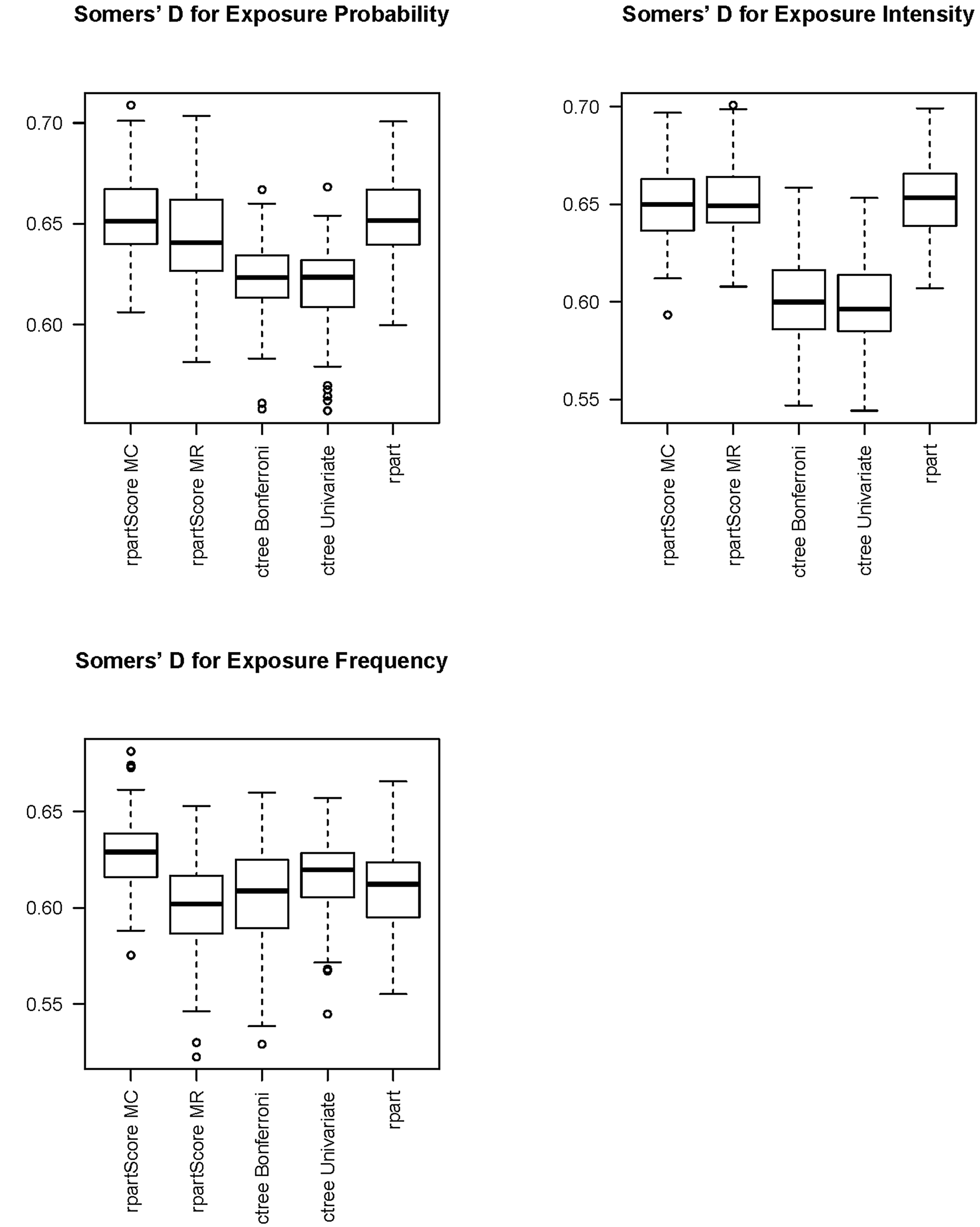

Figure 1 Distribution of Somers' *d* comparing expert estimates and estimates from five CT models for exposure probability, intensity, and frequency in the resampling analysis.

in the median agreement between the rpart method and the rpartScore MC method and ctree Univariate method according to the Wilcoxon tests.

The tendency for rpartScore MC to have fewer large disagreements in exposure estimates than rpart was again evident in the tables of mean agreement between the expert-assigned and model-predicted exposures over the 100 resampled validation data sets for the exposure metrics (see Supplementary Table 4, available at *Annals of Occupational Hygiene* online). For probability, 47% of the jobs assigned a score of 2 by the expert were predicted to be 0 by rpart, whereas rpartScore MC predicted a score of 0 for only 42% of these jobs. For intensity, 29% of the jobs assigned a score of 3 were predicted to be 0 by rpart, but rpartScore MC predicted only 22% of them to be 0. For frequency, rpart predicted 20% of the jobs assigned a score of 3 to have a score of 0, while rpartScore MC predicted 16% of them to be 0.

DISCUSSION

This study compared the predictive ability of multiple CT methods to assess ordinal estimates on a four-category scale for the expert-assigned probability, intensity, and frequency of occupational diesel exhaust exposure estimates in a case–control study. Differences between methods were detected and varied somewhat by the exposure metric. The ordinal rpartScore MC model had a slightly better predictive performance in the resampling analyses than the nominal rpart model for the frequency exposure metric, which we previously found to be the most difficult metric to predict with the nominal tree (Wheeler *et al.*, 2013). We found little difference between the overall predictive performance of the nominal rpart and ordinal rpartScore CT models for the exposure probability and intensity metrics. However, rpartScore MC tended to have fewer large disagreements in exposure estimates compared with rpart across all of the exposure metrics.

Conditional inference trees (ctree models) generally had lower predictive performance than either the nominal or ordinal recursive partitioning (rpart) models in the studied setting. The stopping rules for ctree models are based on *P*-values and tended to result in substantially smaller trees than rpart and rpartScore models, which base the growth and pruning of the tree on prediction error. The ctree Univariate method generally results in a larger tree than does ctree Bonferroni due to a less strict stopping criterion. In this study, ctree Univariate had a better predictive performance for the frequency metric and higher mean agreement across metrics. Given that many exposure scenarios reported within a population-based study are rare, smaller trees would likely miss important exposure distinctions that could be captured by larger trees. This is consistent with early CT research that indicated stopping rules may omit good splits; thus, pruning is preferable to stopping (Breiman *et al.*, 1984).

Overall, our findings provide reassurance that our previous use of nominal CT models using rpart to predict diesel exhaust in this study population (Wheeler *et al.*, 2013) was reasonably robust to the misspecification of the outcome variable. Nominal CTs, which are easy to implement with an R package that includes a graphical user interface (Williams, 2009), provided a reasonable starting point for exploration of predictive performance. However, our findings also suggest that researchers should consider using CT models specifically designed for ordinal outcomes when feasible, especially for situations where the predictive performance of the nominal CT method was less successful. Here, some improvements using ordinal CT methods occurred for the metric for which the nominal CT approach had the lowest agreement (frequency), whereas similar performance between the ordinal and nominal methods was observed for the two metrics that had the highest agreement in the original nominal CT analyses. This suggests that as the practical barriers to using ordinal classification methods dissipate, it becomes preferable to use them in the studied setting. Regardless of the method used, no method had perfect predictive performance. As a result, to avoid the most egregious discrepancies in exposure assignments, CT users would benefit from inspecting the additional predictive information that is provided by each model at a decision rule level (not shown here) to identify where the models had poorer performance to prioritize those jobs for additional expert review.

Differences in performance can be expected between ordinal CTs and nominal CTs when predicting ordinal outcomes due to differences in the node impurity function. One argument against using the nominal Gini impurity function for ordinal data is that it violates one of the ordinal impurity function properties described by Archer and Mas (2009). The violated property is that the impurity function is required to be a non-negative function such that the node impurity is largest when extreme classes in the outcome are equally mixed together (i.e. half of the observations in the no exposure and half in the high exposure groups). In addition, when the predictor variables have monotonic relationships with the ordinal response, the generalized Gini impurity function should lead to better predictive performance. However, when the predictors have a U or J shape with the response, the nominal Gini impurity function may perform well relative to the generalized Gini impurity function because the nominal Gini impurity function is capable of detecting associations between covariates that are not monotonic with the ordinal class (Archer and Mas, 2009). Nevertheless, cross-tabulations of observed versus predicted ordinal response for nominal and ordinal CTs may demonstrate different disagreement patterns and if misclassifications into distant categories are to be penalized more heavily, the generalized Gini impurity function will likely perform better with respect to yielding a higher Somer's *d* (Supplementary Tables 1 and 2, available at *Annals of Occupational Hygiene* online). According to the No Free Lunch theorem (Wolpert and Macready, 1997), no one method will dominate all the others in every situation. Therefore, one cannot make a general statement about which method will perform better when analyzing all future data sets.

A limitation of our work is the use of exposure estimates from one expert assessor as the gold standard with which to evaluate model agreement, as there were no quantitative exposure data or data from more assessors available. The use of one expert is a shortcoming that may have introduced some bias in the estimation of the true exposures. However, the goal of this effort was not to predict truth but rather to replicate the assessments for use in another study. The expert was not involved with the model-building process and hence had no direct influence on the relative performance of the different tree models. Therefore, treating the exposure estimates from one expert as the gold standard is unlikely to explain the differences in relative performance among the tree models.

The recent development of ordinal CT methods somewhat improved our ability to detect relationships between occupational questionnaire responses and expert-assessed exposure estimates for occupational diesel exhaust. However, our findings are based on a single exposure agent in a single study. The effect of CT model selection should be examined for a range of agents, with varying exposure prevalences, numbers of relevant occupational questions, and distributions across exposure categories, to determine whether or not our findings are generalizable to other exposures and studies. Because the best performing model is expected to vary by situation, researchers should consider evaluating several methods to maximize the predictive performance for their data.

SUPPLEMENTARY DATA

Supplementary data can be found at http://annhyg.oxfordjournals.org/.

FUNDING

Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health (Z01 CP10122-19).

ACKNOWLEDGEMENTS

We declare no conflicts of interest.

REFERENCES

Agresti A. (2002) *Categorical data analysis*. 2nd edn. Hoboken, NJ: John Wiley & Sons.

Archer K. (2010) rpartOrdinal: an R package for deriving a classification tree for predicting an ordinal response. *J Stat Softw*; 34: 1–17.

Archer K, Mas V. (2009) Ordinal response prediction using bootstrap aggregation, with application to a high-throughput methylation data set. *Stat Med*; 28: 3597–610.

Behrens T, Mester B, Fritschi L. (2012) Sharing the knowledge gained from occupational cohort studies: a call for action. *Occup Environ Med*; 69: 444–8.

Breiman L, Friedman J, Olshen R *et al.* (1984) *Classification and regression trees*. Pacific Groove, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Carey R, Driscoll T, Peters S *et al.* (2014) Estimated prevalence of exposure to occupational carcinogens in Australia (2011-2012). *Occup Environ Med*; 71: 55–62.

Colt J, Karagas M, Schwenn M *et al.* (2011) Occupation and bladder cancer in a population-based case-control study in Northern New England. *Occup Environ Med*; 68: 239–49.

Friesen MC, Pronk A, Wheeler DC *et al.* (2013) Comparison of algorithm-based estimates of occupational diesel exhaust exposure to those of multiple independent raters in a population-based case-control study. *Ann Occup Hyg*; 57: 470–81.

Galimberti G, Soffritti G, Di Maso M. (2012) Classification trees for ordinal responses in R: the rpartScore package. *J Stat Softw*; 47: 1–25.

Hollander M, Wolfe D. (1999) *Nonparametric statistical methods*. 2nd edn. New York, NY: John Wiley & Sons.

Hothorn T, Hornik K, van de Wiel M *et al*. (2008) Implementing a class of permutation tests: the coin package. *J Stat Softw*; 28: 1–23.

Hothorn T, Hornik K, Zeileis. (2006) Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Stat*; 15: 651–74.

Peters S, Glass D, Milne E *et al*.; the Aus-ALL consortium. (2014) Rule-based exposure assessment versus case-by-case expert assessment using the same information in a community-based study. *Occup Environ Med*; 71: 215–19.

Pronk A, Stewart P, Coble J *et al*. (2012) Comparison of two expert-based assessments of diesel exhaust exposure in a case-control study: programmable decision rules versus expert review of individual jobs. *Occup Environ Med*; 69: 752–8.

R Core Team. (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Therneau T, Atkinson B, Ripley B. (2014) rpart: recursive partitioning and regression trees. R package version 4.1-5. http://CRAN.R-project.org/package=rpart.

Wheeler DC, Burstyn I, Vermeulen R *et al*. (2013) Inside the black box: starting to uncover the underlying decision rules used in a one-by-one expert assessment of occupational exposure in case-control studies. *Occup Environ Med*; 70: 203–10.

Williams GJ. (2009) Rattle: a data mining GUI for R. *R J*; 1: 45–55.

Wolpert D, Macready W. (1997) No free lunch theorems for optimization. *IEEE Trans Evolut Comput*; 1: 67–82.