Ann. Occup. Hyg., 2016, Vol. 60, No. 4, 467–478
doi:10.1093/annhyg/mev095
Advance Access publication 4 January 2016

BOHS
The Chartered Society for
Worker Health Protection

Combining Decision Rules from Classification Tree Models and Expert Assessment to Estimate Occupational Exposure to Diesel Exhaust for a Case-Control Study

Melissa C. Friesen[1,*], David C. Wheeler[2], Roel Vermeulen[3], Sarah J. Locke[1], Dennis D. Zaebst[4], Stella Koutros[1], Anjoeka Pronk[5], Joanne S. Colt[1], Dalsu Baris[1], Margaret R. Karagas[6], Nuria Malats[7], Molly Schwenn[8], Alison Johnson[9], Karla R. Armenti[10], Nathanial Rothman[1], Patricia A. Stewart[11], Manolis Kogevinas[12,13,14] and Debra T. Silverman[1]

1.Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 208952, USA;
2.Department of Biostatistics, Virginia Commonwealth University, Richmond, VA 23298, USA;
3.Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands;
4.Westat, Rockville, MD 20850, USA;
5.TNO, 3700 Zeist, The Netherlands;
6.Geisel School of Medicine at Dartmouth, Hanover, NH 03756, USA;
7.Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Center (CNIO), 28029 Madrid, Spain;
8.Maine Cancer Registry, Augusta, ME 04333-0011, USA;
9.Vermont Cancer Registry, Burlington, VT 05402-0070, USA;
10.New Hampshire Department of Health and Human Services, Division of Public Health Services, Bureau of Public Health Statistics and Informatics, Concord, NH 03301, USA;
11.Stewart Exposure Assessments LLC, Arlington, VA 22207, USA;
12.Centre for Research in Environmental Epidemiology (CREAL), 08003 Barcelona, Spain;
13.CIBER Epidemiología y Salud Pública (CIBERESP), 28029 Barcelona, Spain;
14.IMIM (Hospital del Mar Medical Research Institute), 08003 Barcelona, Spain;
*Author to whom correspondence should be addressed. Tel: +1-240-276-7278; fax: +1-240-276-7835; e-mail: friesenmc@mail.nih.gov

Submitted 8 July 2015; revised 7 December 2015; revised version accepted 10 December 2015.

ABSTRACT

Objectives: To efficiently and reproducibly assess occupational diesel exhaust exposure in a Spanish case-control study, we examined the utility of applying decision rules that had been extracted from expert estimates and questionnaire response patterns using classification tree (CT) models from a similar US study.

Published by Oxford University Press on behalf of the British Occupational Hygiene Society 2016.

Methods: First, previously extracted CT decision rules were used to obtain initial ordinal (0–3) estimates of the probability, intensity, and frequency of occupational exposure to diesel exhaust for the 10 182 jobs reported in a Spanish case-control study of bladder cancer. Second, two experts reviewed the CT estimates for 350 jobs randomly selected from strata based on each CT rule's agreement with the expert ratings in the original study [agreement rate, from 0 (no agreement) to 1 (perfect agreement)]. Their agreement with each other and with the CT estimates was calculated using weighted kappa (κ_w) and guided our choice of jobs for subsequent expert review. Third, an expert review comprised all jobs with lower confidence (low-to-moderate agreement rates or discordant assignments, $n = 931$) and a subset of jobs with a moderate to high CT probability rating and with moderately high agreement rates ($n = 511$). Logistic regression was used to examine the likelihood that an expert provided a different estimate than the CT estimate based on the CT rule agreement rates, the CT ordinal rating, and the availability of a module with diesel-related questions.

Results: Agreement between estimates made by two experts and between estimates made by each of the experts and the CT estimates was very high for jobs with estimates that were determined by rules with high CT agreement rates (κ_w: 0.81–0.90). For jobs with estimates based on rules with lower agreement rates, moderate agreement was observed between the two experts (κ_w: 0.42–0.67) and poor-to-moderate agreement was observed between the experts and the CT estimates (κ_w: 0.09–0.57). In total, the expert review of 1442 jobs changed 156 probability estimates, 128 intensity estimates, and 614 frequency estimates. The expert was more likely to provide a different estimate when the CT rule agreement rate was <0.8, when the CT ordinal ratings were low to moderate, or when a module with diesel questions was available.

Conclusions: Our reliability assessment provided important insight into where to prioritize additional expert review; as a result, only 14% of the jobs underwent expert review, substantially reducing the exposure assessment burden. Overall, we found that we could efficiently, reproducibly, and reliably apply CT decision rules from one study to assess exposure in another study.

KEYWORDS: case-control; diesel exhaust; exposure assessment methodology; occupational exposure; statistical learning

INTRODUCTION

Occupational exposure assessment in case-control studies that include detailed occupational questionnaires, such as exposure-oriented modules targeting specific jobs (e.g. truck drivers) or industries (e.g. railroad industry), has become more efficient and reproducible with the development of methods that systematically apply programmable exposure decision rules that explicitly link questionnaire response patterns to exposure decisions (Fritschi *et al.*, 2009; Behrens *et al.*, 2012; Pronk *et al.*, 2012; Friesen *et al.*, 2013; Wheeler *et al.*, 2013; Carey *et al.*, 2014; Peters *et al.*, 2014). In addition, we recently demonstrated that we could extract previously made decision rules using nominal classification tree (CT) models that identified patterns between questionnaire responses and expert-based exposure estimates (Wheeler *et al.*, 2013, 2015). CT models predict an outcome using a sequential splitting approach, where each sequential split is made using the most predictive variable, until the model meets user-specified stopping criteria with which to identify an outcome for each branch of the model. We found that predicted estimates from nominal CT models replicated 83–92% of an expert's ordinal assignments of probability, intensity, and frequency of occupational diesel exhaust exposure for jobs reported in the New England Bladder Cancer Study (NEBCS) (Wheeler *et al.* 2013). In addition to providing the best-fitting exposure assignment, the CT models provided for each rule the agreement rate between the model prediction and the expert's rating. This agreement rate provides a potential metric that can be used to efficiently identify a smaller set of jobs with questionnaire response patterns that might require further expert-based review. Thus far, our evaluations of CT models have focused on their ability to replicate expert decisions within a validation set from the same study. The utility of applying the CT decision rules to aid exposure assessment in other studies has not yet been evaluated.

The purpose of the current study was to gain insights into the applicability of CT decision rules extracted from one study to estimate exposure in another study. Here, we describe our application of the extracted CT decision rules for occupational diesel exhaust exposure from NEBCS, which were reported in the study by Wheeler *et al.* (2013), to the Spanish Bladder Cancer Study (SBCS). This second study used the NEBCS occupational questionnaires after translation into Spanish (Samanic *et al.*, 2008). Our aims were to leverage the previous decision rules, to conduct a reliability assessment for a subset of jobs to develop criteria for identifying jobs requiring further expert review, and to efficiently incorporate expert review to obtain reliable estimates of occupational exposure to diesel exhaust for future SBCS epidemiologic analyses.

METHODS

NEBCS exposure metrics and CT decision rules

In NEBCS, estimates of occupational diesel exhaust exposure were assigned from a job-by-job expert review of the 14 983 jobs reported by NEBCS participants (Pronk *et al.*, 2012). Based on information provided in the occupational histories and in exposure-oriented modules targeting specific jobs or industries and a review of the literature (Pronk *et al.*, 2009), the expert assigned ordinal estimates of the probability of exposure and continuous measures of the intensity (μg m^{-3} respirable elemental carbon, REC) and frequency (hours per week) of exposure. Probability was assessed as the estimated proportion of workers likely exposed to diesel exhaust based on the reported information, with cutpoints of <5% (0), 5–49% (1), 50–79% (2), and ≥80% (3). Intensity was assessed on a continuous scale as the estimated average personal REC concentration (μg m^{-3}) during diesel-exposed tasks, and categorized for the CT models using cut points of <0.25 (0), 0.25 to <5 (1), 5 to <20 (2), and ≥20 μg m^{-3} REC (3), which represented natural groupings in the continuous metric. Frequency was assessed on a continuous scale as the estimated average number of hours per week exposed to diesel exhaust, and categorized (based on observed groupings in the continuous metric) for the CT models using cut points of <0.25 (0), 0.25 to <8 (1), 8 to <20 (2), and ≥20 hours per week (3).

The CT decision rules derived from NEBCS-based nominal CT models were reported in the online supplementary materials of Wheeler *et al.* (2013). These decision rules do not necessarily represent the expert's decision process, but instead represent the questionnaire response patterns that best (agnostically) replicated an expert's exposure decision. Briefly, the CT models used a training data set of 10 488 (70%) jobs to develop nominal CT models that identified patterns between variables extracted from the ordinal expert estimates of probability, intensity and frequency, and diesel-related variables derived from the occupational questionnaires. The occupational questionnaire variable set comprised 498 variables extracted from the occupational histories and 223 variables extracted from the exposure-oriented modules. In addition, we developed a CT model that explained a dichotomous probability estimate derived from the ordinal probability metric (0 = <50%; 1 = ≥50% probability). In total, we extracted four sets of decision rules, one each for ordinal probability, ordinal intensity, ordinal frequency, and dichotomous probability. The extracted CT decision rules used 106 (15%) of the diesel-related variables extracted from the occupational histories and modules.

The CT model provides the rating, the CT agreement rate with the expert assignment, and the set of conditions defined by the occupational variables that apply for each decision rule. The CT agreement rate is the proportion of the observations meeting the specified rule criteria that fall into each expert-assigned rating category, on a continuous scale from 0 to 1. The following is an example of a CT decision rule to estimate probability of diesel exposure:

Rule #1: rating = 3; probabilities: 0.007 0.000 0.013 0.980

Worked near or smelled exhaust = yes

Equipment powered by diesel = yes

In the above example, the first row indicates that the assigned rating was 3 and that the CT agreement rates for ratings 0, 1, 2, and 3 were estimated as 0.007, 0.000, 0.013, and 0.980, respectively. The highest CT agreement rate was used to assign the CT rating. Here, the highest agreement rate was for rating category 3; its agreement rate of 0.980 indicates that 98% of the expert's ratings for jobs with the rule's criteria had a rating of 3. In this example, two variables defined the conditions for this rule, which were jobs that the participant reported working near or smelling exhaust (second row) 'and' indicated working with equipment powered by diesel engines (third row).

SBCS population

The SBCS population has been previously described (Samanic *et al.*, 2008). Briefly, SBCS included 1219 cases with urothelial carcinoma of the urinary bladder and 1271 controls selected from 18 hospitals in Spain between June 1998 and September 2000. Each participant completed a lifetime occupational history questionnaire as part of an interview capturing information on smoking history, medical history, demographics, and other risk factors. The occupational history comprised open-ended questions asking about the job title, name and location of the employer, type of service or product provided, year started and stopped, work frequency, activities and tasks, the tools and equipment used, and the chemicals and materials handled. For each job, participants were also asked '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'. The responses in the occupational histories and to these two engine and engine exhaust questions triggered modules that asked more detailed exposure-oriented questions for 62% of the 10 182 SBCS jobs. Modules with diesel-related questions were completed for 33% of the SBCS jobs.

Assignment of diesel exhaust exposure estimates to SBCS

SBCS data re-coding

To apply the NEBCS CT decision rules to SBCS jobs, some data re-coding of SBCS occupational variables was necessary. Re-coding efforts in SBCS were restricted to questionnaire responses that were associated with the 106 questionnaire variables included in the NEBCS CT decision rules from nominal CT models. Other diesel-related questions/variables were not used in the CT decision rules and were not re-coded.

CT decision rule application

The NEBCS CT decision rules were converted to Stata commands (Stata S.E. v.11.2, StataCorp LP, College Station, TX, USA) that assigned the following to each SBCS job for each of the four exposure metrics: (i) the CT decision rule number; (ii) the CT rule's agreement rate for categories 0, 1, 2, and 3, respectively, based on the NEBCS expert estimates; (iii) the CT rule's estimate with the highest CT agreement rate (assigned CT estimate); and (iv) the CT rule's agreement rate for the assigned CT estimate (highest CT agreement rate).

Reliability assessment

Our prior evaluations found that the NEBCS nominal CT models replicated 83–92% of the expert-assigned estimates in the validation set (Wheeler *et al.*, 2013). However, the agreement varied by rating, with low agreement for some exposure ratings, suggesting that an expert review stage would be needed for at least a subset. To identify where to focus expert review, we first conducted a reliability assessment where two industrial hygienists (S.J.L. and D.D.Z.) independently reviewed the CT estimates for a subset of SBCS jobs selected as described below and, where appropriate, provided an updated estimate. We hypothesized that changes were more likely occur when the CT agreement rate was low, when there was discordance between metrics (e.g. one metric indicated exposed, another unexposed), and for the frequency metric, which had the lowest agreement in previous NEBCS comparisons. Thus, we categorized all SBCS jobs into one of four strata using the following criteria:

A. *High confidence*: jobs assigned estimates based on rules that had highest CT agreement rates ≥0.7 for the ordinal probability rating and ≥0.6 for the ordinal intensity and frequency ratings, and that had no discrepancy between the ordinal and dichotomous CT probability ratings;
B. *Discordant probability ratings*: jobs with a discrepancy between the ordinal and dichotomous CT probability ratings (i.e. either ordinal rating of 2 or 3 and dichotomous rating of 0 *or* ordinal rating of 0 or 1 and dichotomous rating of 1), but with CT agreement rates that met the criteria for stratum 'A';
C. *Low frequency confidence*: jobs with CT agreement rates ≥0.7 for the ordinal probability rating, ≥0.6 for the ordinal intensity rating, but <0.6 for the ordinal frequency rating;
D. *Low probability/intensity confidence*: jobs with CT agreement rates <0.7 for the ordinal probability rating and/or <0.6 for the ordinal intensity rating.

From stratum A, we randomly selected 50 jobs with a CT ordinal probability rating of 0 or 1 and 50 jobs with an ordinal probability rating of 2 or 3. From stratum B, which had the fewest number of jobs, we randomly selected 50 jobs; from strata C and D, we randomly selected 100 jobs. The actual distribution of jobs across strata shown in Table 1 differed slightly from these intentions due to a correction in variable coding that occurred after the reliability subset was identified. For each stratum and for each exposure metric, we evaluated the agreement between the two experts' estimates and between each expert's estimates and the CT estimates. Because the above definitions were arbitrary, within each stratum we also evaluated agreement within subsets of jobs, such as jobs with a nonzero probability rating versus jobs with a probability ratings of 0 or those with highest CT agreement rates ≥0.6 to <0.7, ≥0.7 to <0.8, and ≥0.8. Agreement was calculated using the proportion of assignments with exact agreement and weighted kappa with quadratic weights. These agreement metrics were used to determine whether the magnitude of agreement between the CT estimates and each of the experts was similar in magnitude to the agreement between the two experts, with the aim to focus additional expert review on jobs where lower agreement was observed using the CT estimates.

Additional expert review

The results of the reliability assessment (described in the Results section) guided our second round of job-by-job expert review. Because of the poor-to-moderate agreement between the CT estimates and each expert for strata B, C, and D, one of the experts (S.J.L.) reviewed all jobs in these strata. Because of excellent agreement for stratum A and to prioritize specificity rather than sensitivity in capturing exposure status, the review of jobs in stratum A was limited to those that had a CT ordinal probability rating of 2 or 3 (≥50% probability) 'and' that met one of the following conditions: (i) a probability CT agreement rate of 0.7 to <0.8 or (ii) intensity and/or frequency CT agreement rate of 0.6 to <0.8. The expert involved in this analysis had extensively reviewed the previously developed decision rules that were the basis of the NEBCS expert review (Pronk *et al.*, 2012). The expert reviewed the questionnaire responses and the corresponding CT estimates and, where the CT estimate was inconsistent with the previously articulated rules, provided a different estimate. The number of times the expert estimate was different than the CT estimate was calculated for each metric.

Identifying characteristics of expert changes

The strata and conditions chosen for expert review were based on hypotheses of where an expert would more likely provide an estimate that differed from the CT estimate (hereafter, a 'changed' estimate), but these hypotheses have not been previously tested. To gain insight into whether our criteria for expert review were appropriate, we used all jobs that underwent expert review to evaluate whether the expert was more likely to change the CT estimate at varying CT agreement rates (categorized as <0.7, 0.7 to ≤0.8, >0.8 to ≤0.9, and >0.9), at varying CT ordinal rating categories (0–3), and whether a module with diesel-related questions was available for that job (yes versus no). We developed separate weighted logistic regression models for each exposure metric and for two definitions of change in assignment (definition A: ≥1 versus no category change; definition B: ≥2 versus <2 category change). Jobs in stratum A were weighted by the inverse probability of being selected for expert review; jobs in the other strata were given a weight = 1, because all jobs in strata B, C, and D were reviewed.

RESULTS

After application of the CT decision rules, 91% of the 10 182 SBCS jobs were categorized into the 'high confidence' stratum and 9% were distributed across strata identifying discordance and/or lower confidence (Table 1). In the high confidence stratum, 85% had a CT ordinal probability estimate of 0 (<5% probability); jobs in the other strata were, by definition, much less likely to have a CT ordinal probability rating of 0 (0–52%). The reliability and subsequent expert review subsets demonstrated our oversampling and prioritization of jobs with lower confidence and CT probability rating >0. Overall, 1442 (14%) jobs underwent expert review (including jobs reviewed in the reliability assessment). Of these, only 16% had an initial CT probability rating of 0.

In the reliability assessment, the agreement between two experts and between each expert and the CT estimates varied by exposure metric and by strata defined by the CT confidence and discordance

Table 1. Distribution of SBCS jobs by strata defined by CT confidence, for all jobs, for the reliability subset, and for the jobs that underwent expert review

Strata	All Jobs			Reliability subset			Jobs that underwent expert review[a]		
	N jobs	% all jobs	% CT ordinal prob. =0 in strata	N jobs selected	% jobs in strata selected	% CT ordinal prob. = 0 in strata	N jobs with expert review	% jobs in strata selected for expert review	% CT ordinal prob. = 0 in strata
All jobs	10 182	100%	79%	350	3%	23%	1442	14%	16%
By strata defined by confidence in CT estimate[b]									
A. High conf.	9251	91%	85%	123	1%	40%	511	6%	25%
B. Discordant prob.	143	1%	52%	48	34%	58%	143	100%	52%
C. Low frequency conf.	436	4%	0%	82	19%	0%	436	100%	0%
D. Low prob./intensity conf.	352	3%	9%	97	28%	5%	352	100%	9%

conf., confidence; prob., probability rating.

[a] Includes reliability subset plus the additional jobs identified for expert review (described in section on Additional Expert Review).

[b] Definitions of strata are provided in section on Reliability Assessment.

between metrics (Table 2). In stratum A (high confidence), very high agreement for all three metrics was observed between the two experts and between each expert and the CT estimates (κ_w range 0.81–0.90). In this stratum, the agreement for the intensity and frequency estimates were somewhat lower when restricted to jobs with CT probability rating >0 than for all jobs assessed in this strata, but the magnitude was similar across the three sets of comparisons. For the probability metric, the proportion agreement between the CT estimates and each expert was 5–15% lower than between the two experts, both overall and for jobs with a CT probability rating >0. In stratum B (discordant probability estimates), moderate agreement was observed between the two experts (κ_w range 0.42–0.66) and poor-to-moderate agreement was observed between the experts and CT estimates (κ_w range 0.24–0.57). In strata C and D (low frequency confidence and low probability/intensity confidence, respectively), the agreement between the two experts was moderate (κ_w range 0.47–0.68) and consistently higher than between the experts and the CT model (κ_w range 0.09–0.52). The same patterns were observed for all 1442 SBCS jobs included in the expert review (data not shown).

Table 3 shows the overall distribution of the CT estimates before and after the two sets of expert review, as well as the number of estimates changed in the expert review of 1442 jobs. The overall distribution remained approximately the same, with the expert review classifying a slightly greater number of jobs to a rating of 3 for intensity (1.2 versus 0.8%) and frequency (7.9%

Table 2. Agreement among ordinal occupational diesel exposure estimates from two experts and from CT models for 350 jobs, by exposure metric and strata defined by CT models' assessment confidence

Strata[a]	Comparison group	Probability		Intensity		Frequency	
		% Agree.	κ_w	% Agree.	κ_w	% Agree.	κ_w
A. High conf.	Expert 1 versus 2	90.2	0.86	78.0	0.84	72.4	0.84
	CT prob >0	90.5	0.56	71.6	0.72	62.2	0.71
	Expert 1 versus CT	83.7	0.82	88.6	0.85	69.9	0.84
	CT prob >0	79.7	[b]	74.3	0.60	73.0	0.79
	Expert 2 versus CT	85.4	0.83	80.5	0.81	81.3	0.90
	CT prob >0	75.7	[b]	85.1	0.71	54.1	0.69
B. Discordant prob.	Expert 1 versus 2	68.8	0.42	77.1	0.51	72.9	0.66
	Expert 1 versus CT	83.3	0.46	89.6	0.57	87.5	0.43
	Expert 2 versus CT	75.0	0.32	72.9	0.24	72.9	0.39
C. Low frequency conf.	Expert 1 versus 2	79.3	0.50	69.5	0.57	67.1	0.67
	Expert 1 versus CT	82.9	0.24	80.5	0.52	34.1	0.17
	Expert 2 versus CT	85.4	0.22	67.1	0.23	40.2	0.22
D. Low prob./ intensity conf.	Expert 1 versus 2	68.0	0.63	64.9	0.47	56.7	0.68
	Expert 1 versus CT	61.9	0.09	67.0	0.34	62.9	0.44
	Expert 2 versus CT	63.9	0.20	64.9	0.27	51.5	0.28

κ_w, weighted kappa; Agree., agreement; CT, classification tree; conf., confidence; prob., probability rating.

[a] Definitions of strata are provided in section on Reliability Assessment.

[b] Restricting the jobs to CT probability >0 turns the CT probability metric into a 3 category rather than 4 category assessment. As a result, weighted kappa's between the CT estimate and the expert estimate cannot be calculated.

Table 3. Prevalence of ratings and number of changed estimates after expert review of 1442 jobs

Exposure metric Ordinal rating		CT estimate	Estimate after expert review[a]	*N* estimates changed[b] (%)
		N jobs (%)	*N* jobs (%)	
Probability				
	0	7942 (78)	7983 (78)	41 (0.5)
	1	651 (6.4)	573 (5.6)	78 (12)
	2	170 (1.7)	197 (1.9)	27 (16)
	3	1419 (14)	1429 (14)	10 (0.7)
	Overall	10 182 (100)		156 (1.5)
Intensity				
	0	7701 (76)	7637 (75)	64 (0.8)
	1	1968 (19)	1983 (19)	15 (0.8)
	2	432 (4.2)	445 (4.4)	13 (3.0)
	3	81 (0.8)	117 (1.2)	36 (44)
	Overall	10 182 (100)		128 (1.3)
Frequency				
	0	8133 (80)	7826 (77)	307 (3.8)
	1	872 (8.6)	922 (9.1)	50 (5.7)
	2	563 (5.5)	635 (6.2)	72 (13)
	3	614 (6.0)	799 (7.9)	185 (30)
	Overall	10 182 (100)		614 (6.0)

[a] Includes 1442 jobs within the reliability study and the second round of expert review.
[b] For jobs in the reliability study, the number of estimates changed were based on the estimates of the same expert that conducted the second round of expert review.

versus 6.0%) compared to the CT estimate. The expert review changed 156 (1.5%), 128 (1.3%), and 614 (6.0%) of the CT estimates for the probability, intensity, and frequency metrics, respectively. The probability estimates were changed by ≥1 category for 20% and ≥2 categories for 15% of the 1,442 jobs, the intensity estimates were changed by ≥1 category for 21% and ≥2 categories for 5% of the jobs, and the frequency estimates were changed by ≥1 category for 46% and ≥2 categories for 23% of the jobs (data not shown).

Results from logistic regression models generally showed similar patterns across all exposure metrics and both definitions of change (Table 4). Expert changes in estimates were most likely to occur when the CT agreement rate was ≤0.7 (odds ratio, OR range: 2.6–27) or >0.7 to ≤0.8 (OR range 1.7–6.1). For frequency, however, the likelihood of a change remained high (OR 5.0), even when CT agreement rates were between >0.8 and ≤0.9. Changes in estimates were generally less frequent when the CT estimate was 3 (OR range 0.15–0.51) or 0 (reference, OR = 1). Patterns for ratings of 1 and 2 varied by metric. For probability, compared to a rating of 0, CT ratings of 1 were the most likely to be changed ≥1 category (OR 9) or ≥2 categories (OR 38). In contrast, probability CT ratings of 2 had a 2.6 times increased likelihood

Table 4. Characteristics predicting the relative likelihood (odds ratio) that an expert changed the CT ordinal ratings by ≥1 category and ≥2 categories based on logistic regression models

	Probability					Intensity					Frequency				
	% of jobs[a]	≥1 category change ($n = 293$)		≥2 category change ($n = 215$)		% of jobs	≥1 category change ($n = 310$)		≥2 category change ($n = 68$)		% of jobs	≥1 category change ($n = 663$)		≥2 category change ($n = 328$)	
		OR	CI	OR	CI		OR	CI	OR	CI		OR	CI	OR	CI
CT agreement rate															
>0.9	32	Ref.		Ref.		18	Ref.		Ref.		11	Ref.		Ref.	
0.8 to ≤0.9	12	0.23	0.06–0.8	0.30	0.08–1.2	19	0.7	0.14–3.0	0.10	0.02–0.46	47	5.0	1.3–19	0.41	0.07–2.23
0.7 to ≤0.8	7	1.7	0.62–4.8	1.9	0.68–5.6	15	6.1	1.3–28	0.38	0.05–2.6	28	5.9	1.0–8.0	2.0	0.52–8.0
≤0.7	50	2.6	0.81–8.5	3.1	0.93–10	48	7.2	1.9–27	0.69	0.10–4.6	14	27	7.7–97	16	3.9–69
CT rating[b]															
0	16	Ref.		Ref.		42	Ref.		Ref.		25	Ref.		Ref.	
1	9	38	13–114	9.1	3.1–27	23	2.8	0.98–8.0	1.4	0.39–5.3	58	0.23	0.07–0.82	0.19	0.05–0.74
2	10	2.6	0.93–7.2	0.53	0.17–1.6	16	2.4	0.81–7.3	33.4	10–109	15	5.8	1.8–18	2.9	0.73–12
3	64	0.15	0.06–0.37	0.12	0.05–0.31	18	0.51	0.09–2.8	—		1	0.25	0.07–0.90	0.20	0.04–1.0
Diesel module															
No	45	Ref.		Ref.		45	Ref.		Ref.		45	Ref.		Ref.	
Yes	55	2.0	0.85–4.7	2.3	1.0–5.5	55	1.7	0.78–3.7	0.90	0.22–3.6	55	2.1	1.1–4.2	2.1	0.91–5.0

CI, confidence interval; OR, odds ratio; Ref., reference group.

[a] Jobs in the reliability assessment or in the subsequent expert review ($n = 1442$).

[b] CT rating for the metric being assessed.

of being changed ≥1 category; however, this change was usually a one category change, because ratings of 2 were less likely to have a change of ≥2 categories (i.e. from exposed to unexposed; OR 0.5). For intensity, compared to a rating of 0, CT ratings of 1 and 2 had a 2.4–2.8 times increased likelihood of being changed ≥1 category; results were unstable for assessments of 2+ category changes, which were infrequent ($n = 68$). For frequency, compared to a rating of 0, ratings of 1 were less likely to be changed (OR ~0.2), but ratings of 2 were 5.8 and 2.9 times more likely to be changed ≥1 category and ≥2 categories, respectively. The availability of a diesel-related module increased the likelihood that the expert changed an estimate by approximately 2-fold (OR range 1.7–2.1 for ≥1 category; 0.9–2.3 for ≥2 category) compared to when only the occupational history was available.

DISCUSSION

This study demonstrated the feasibility of applying CT decision rules from one study to another study that used the same (translated) occupational questions to estimate occupational diesel exhaust exposure. Our approach incorporated expert review only when the CT models' estimates were less confident or where the CT estimates of different metrics were discordant, resulting in only 14% of the jobs reviewed one-by-one. Of these, only 156 of the probability estimates (1.5% of the jobs overall), 128 of the intensity estimates (1.3% overall), and 614 of the frequency estimates (6.0% overall) were changed. Thus, our results suggest that the CT models usually provided reliable probability and intensity estimates that replicated expert job-by-job estimates for the majority of jobs. Frequency remained challenging to reliably assign using CT rules, which was consistent with our previous evaluations (Wheeler *et al.*, 2013, 2015). Moreover, we demonstrated that the likelihood that an expert changed an estimate was closely associated with lower CT agreement rates, the middle exposure categories, and the availability of additional information from an exposure-oriented module. These findings can be used in subsequent applications of the rules to prioritize expert review for a more efficient exposure assessment.

A key feature of the application of the CT rules to SBCS was the use of a reliability assessment to evaluate whether our initial hypotheses of where to focus expert review was appropriate. Here, because this was the first application of CT rules outside the original study and because we wanted to identify features of the CT decision rules that might indicate when an expert review was needed, we kept our review inclusion criteria broad and purposely reviewed more jobs than likely necessary. The high agreement in our 'high confidence' stratum (A) was expected because of the high prevalence of unexposed jobs (85% overall; 23% in stratum A), which prior evaluations found the easiest to identify (Pronk *et al.*, 2012; Friesen *et al.*, 2013; Wheeler *et al.*, 2013, 2015). As hypothesized, poor-to-moderate agreement between CT estimates and expert estimates was observed in the strata defined by greater uncertainty. Moreover, the agreement between two experts decreased from high in our 'high confidence' stratum to only moderate or moderately high in these lower confidence strata, providing a reminder that many jobs remain difficult to reliably assess using expert-based approaches. The assessments relied on the information provided by the participants and those responses may be interpreted and assessed differently by experts with differing backgrounds and expertise with diesel exhaust exposure.

Our evaluations demonstrated that the CT agreement rate was a good predictor of where an expert was likely to change a CT estimate, but we suspect that the observed patterns based on CT rating and presence of an exposure-oriented module may vary by agent and study. Here, the patterns where the expert changed CT estimates were consistent with our prior evaluations for diesel exhaust exposure estimates in NEBCS. For example, in NEBCS we demonstrated that CT estimates and expert estimates had high agreement for ratings of 0 and 3 and poor-to-moderate agreement for the middle categories (Wheeler *et al.*, 2013). The challenge with the middle categories also parallels the greater disagreements between experts for those categories (Friesen *et al.*, 2013). The greater likelihood of a change in an assessment when a diesel-relevant module was available may reflect the exposure prevalence differences in the two groups, where the proportion of jobs with probability rating ≥2 was 8% for jobs without a diesel-related module and 31% for jobs with a diesel-related module. It may also reflect the occurrence of rarely reported exposure scenarios that were too infrequent to be captured in the CT models' decision rules, which depends on the number of variables available and CT model specifications (e.g. the minimum number of observations needed to identify a 'split' in the tree or to define a 'leaf' that indicates an exposure decision).

Post hoc discussions with the industrial hygienist conducting the expert review and a review of her written reasons for revising the CT estimates identified several broad patterns for the changes. First, SBCS included more jobs pre-1960 (48% of SBCS jobs started pre-1960), when diesel engine use was rarer, than did the NEBCS (31%). Thus, the CT decision rules were unable to fully capture time period changes in expert estimates for the earliest time periods. Second, the contextual occupation information provided by SBCS participants for jobs in agricultural and mining industries indicated much smaller operations, more manual work, and more work that used horses, than observed for NEBCS' jobs. Third, frequency remained difficult to assess using formal rules because this metric often required aggregating time spent on multiple tasks with potential diesel exposure. Identifying when tasks occurred concurrently or separately was often difficult to disentangle even during expert review and often required additional contextual information from responses to other questions.

Strengths and limitations

This study's strengths include the availability of CT decision rules for occupational diesel exhaust exposure extracted from over 10 000 jobs (Wheeler *et al.*, 2013), inclusion of two experts involved in previous evaluations of the reliability of diesel exposure estimates in NEBCS such that the underlying rules remained consistent (Friesen *et al.*, 2013), and the availability of a second study that used the same (translated) occupational questions.

The major limitation, and the biggest challenge for any exposure assessment in population-based studies, is the lack of gold standards with which to compare exposure assessment approaches for even a subset of jobs. The good agreement observed here between CT estimates and expert estimates in our high confidence strata indicated only that the CT estimates can reliably replicate an expert's estimates and provided a measure of how good that replication is. We are unable to evaluate whether one approach is more valid than the other.

A second limitation is that the CT estimates that were changed during expert review were based on the assessments of only one expert. Previous studies have shown that the validity of the estimates may increase with the use of teams of experts and aggregating their ratings (de Cock *et al.*, 1996; Semple *et al.*, 2001; Steinsvag *et al.*, 2007; Friesen *et al.*, 2011). In practice, multiple experts are expensive and often infeasible. We used a single expert calibrated to the original NEBCS decision rationales because we wanted our evaluations to measure whether the CT estimates replicated the original rationales (Pronk *et al.*, 2012) rather than also reflecting differing opinions about the decision rules.

There are several additional limitations. We were unable to refine the CT rules for country-specific differences because of limited information to inform those refinements. Spanish data from 1965 indicate that diesel fueled over 90% of industrial trucks, over 80% of buses, and over 40% of trucks and vans, with increasing prevalence over time. US data for this same period suggest that only ~40% of industrial trucks and 5–10% of trucks and vans were fueled by diesel, suggesting that the probability, frequency, and intensity of exposure in traffic-exposed jobs may have been potentially higher in Spain than in the USA. In addition, the two countries likely had industry-specific differences in the prevalence of use of diesel powered vehicles and equipment. Thus, decision rules developed for one study population must be evaluated as to their applicability to a second population to ensure that the exposure situations are similar and revised to reflect known differences. This requirement, however, is similar to that of applying a job-exposure matrix developed for one population to other populations (Lavoue *et al.*, 2012).

An additional limitation is that the CT decision rules were based on nominal rather than ordinal CT models. We recently found that ordinal CT models had a similar ability to replicate an expert's estimates to the nominal models but their use reduced the number of ≥2 category errors and slightly improved the replication of the frequency estimates (Wheeler *et al.*, 2015). Thus, the use of the ordinal CT models would be expected to reduce the number of jobs with ≥2 category changes, but not necessarily reduce the number of ≥1 category changes. However, because we were able to use the CT agreement rates as a measure of confidence to identify when the assigned rating is less certain. We expect that these more egregious errors would have low CT agreement rates and be included in the expert review. Future work could include comparing the number of changes for nominal versus ordinal CT models, but this approach would require additional recoding efforts in SBCS to include variables predictive in the ordinal but not nominal CT models.

Finally, these evaluations were limited to a single agent—diesel exhaust—that had a moderate prevalence of exposure and that was a focus of the occupational questionnaires. It is unknown whether we would be as successful in extracting and applying CT decision rules for other agents that may have lower exposure prevalence or less coverage in the questionnaires. However, the use of reliability assessments in informative subsets can be used in other studies to provide study- and agent-specific insights into where to focus expert review.

CONCLUSION

We found that, to replicate a job-by-job expert review, we could efficiently, reproducibly, and reliably apply CT decision rules from one study to assess exposure for the majority of jobs in another study with detailed occupational information, with only a fraction (14%) requiring review by an expert. In subsequent development and application of CT decision rules, a reliability assessment may be a helpful tool to identify where to focus expert review. Future work will be needed to evaluate whether CT decision rules for occupational diesel exhaust exposure could be applied in other studies with similar but not identical questions. In the absence of gold standards of exposure, developing methods to efficiently assign reproducible exposure decisions in population-based studies improves our ability to evaluate that exposure in more studies and to evaluate the associations' robustness to exposure assessment decisions.

ACKNOWLEDGEMENTS

This study was funded by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health. We also thank Dr M. García-Closas for her involvement in the Spanish Bladder Cancer Study. The authors report no conflicts of interest.

REFERENCES

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.

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

de Cock J, Kromhout H, Heederik D *et al.* (1996) Experts' subjective assessment of pesticide exposure in fruit growing. *Scand J Work Environ Health*; 22: 425–32.

Friesen MC, Coble JB, Katki HA *et al.* (2011) Validity and reliability of exposure assessors' ratings of exposure intensity by type of occupational questionnaire and type of rater. *Ann Occup Hyg*; 55: 601–11.

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 casecontrol study. *Ann Occup Hyg*; 57: 470–81.

Fritschi L, Friesen MC, Glass D *et al.* (2009) Occideas: Retrospective occupational exposure assessment in community-based studies made easier. *J Environ Public Health*; 2009: 957023.

Lavoue J, Pintos J, Van Tongeren M *et al.* (2012) Comparison of exposure estimates in the Finnish job-exposure matrix finjem with a JEM derived from expert assessments performed in Montreal. *Occup Environ Med*; 69: 465–71.

Peters S, Glass DC, Milne E *et al.* (2014) Rule-based exposure assessment versus case-by-case expert assessment using the same information in a community-based study. *Occup Environ Med*; 71: 215–9.

Pronk A, Coble J, Stewart PA. (2009) Occupational exposure to diesel engine exhaust: a literature review. *J Expo Sci Environ Epidemiol*; 19: 443–57.

Pronk A, Stewart PA, Coble JB *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.

Samanic CM, Kogevinas M, Silverman DT *et al.* (2008) Occupation and bladder cancer in a hospital-based case-control study in Spain. *Occup Environ Med*; 65: 347–53.

Semple SE, Proud LA, Tannahill SN *et al.* (2001) A training exercise in subjectively estimating inhalation exposures. *Scand J Work Environ Health*; 27: 395–401.

Steinsvag K, Bratveit M, Moen BE *et al.* (2007) Inter-rater agreement in the assessment of exposure to carcinogens in the offshore petroleum industry. *Occup Environ Med*; 64: 582–88.

Wheeler DC, Archer KJ, Burstyn I *et al.* (2015) Comparison of ordinal and nominal classification trees to predict ordinal expert-based occupational exposure estimates in a case-control study. *Ann Occup Hyg*; 59: 324–35.

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.