Ann. Occup. Hyg., 2015, Vol. 59, No. 4, 455–466
doi:10.1093/annhyg/meu101
Advance Access publication 3 December 2014

Using Hierarchical Cluster Models to Systematically Identify Groups of Jobs With Similar Occupational Questionnaire Response Patterns to Assist Rule-Based Expert Exposure Assessment in Population-Based Studies

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

1.Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA
2.Biostatistics, Group Health Research Institute, Seattle, WA 98101-1448, USA
3.Department of Biostatistics, Virginia Commonwealth University, Richmond, VA 23298, USA
4.Department of Environmental and Occupational Health, Drexel University, Philadelphia, PA 19104, USA
5.Utrecht University, Utrecht, The Netherlands
6.TNO, Utrecht, The Netherlands
7.Geisel School of Medicine at Dartmouth, Hanover, NH 03756, USA
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.Biostatistics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA
*Author to whom correspondence should be addressed. Tel: 301-594-7485; fax: 240-276-7835; e-mail: friesenmc@mail.nih.gov
Submitted 7 March 2014; revised 21 October 2014; revised version accepted 27 October 2014.

ABSTRACT

Objectives: Rule-based expert exposure assessment based on questionnaire response patterns in population-based studies improves the transparency of the decisions. The number of unique response patterns, however, can be nearly equal to the number of jobs. An expert may reduce the number of patterns that need assessment using expert opinion, but each expert may identify different patterns of responses that identify an exposure scenario. Here, hierarchical clustering methods are proposed as a systematic data reduction step to reproducibly identify similar questionnaire response patterns prior to obtaining expert estimates. As a proof-of-concept, we used hierarchical clustering methods to identify groups of jobs (clusters) with similar responses to diesel exhaust-related questions and then evaluated whether the jobs within a cluster had similar (previously assessed) estimates of occupational diesel exhaust exposure.

Methods: Using the New England Bladder Cancer Study as a case study, we applied hierarchical cluster models to the diesel-related variables extracted from the occupational history and job- and industry-specific questionnaires (modules). Cluster models were separately developed for two subsets: (i) 5395 jobs with ≥1 variable extracted from the occupational history indicating a potential diesel exposure scenario, but without a module with diesel-related questions; and (ii) 5929 jobs with both occupational history and module responses to diesel-relevant questions. For each subset, we varied the numbers of clusters extracted from the cluster tree developed for each model from 100 to 1000 groups of jobs. Using previously made estimates of the probability (ordinal), intensity (μg m^{-3} respirable elemental carbon), and frequency (hours per week) of occupational exposure to diesel exhaust, we examined the similarity of the exposure estimates for jobs within the same cluster in two ways. First, the clusters' homogeneity (defined as >75% with the same estimate) was examined compared to a dichotomized probability estimate (<5 versus ≥5%; <50 versus ≥50%). Second, for the ordinal probability metric and continuous intensity and frequency metrics, we calculated the intraclass correlation coefficients (ICCs) between each job's estimate and the mean estimate for all jobs within the cluster.

Results: Within-cluster homogeneity increased when more clusters were used. For example, ≥80% of the clusters were homogeneous when 500 clusters were used. Similarly, ICCs were generally above 0.7 when ≥200 clusters were used, indicating minimal within-cluster variability. The most within-cluster variability was observed for the frequency metric (ICCs from 0.4 to 0.8). We estimated that using an expert to assign exposure at the cluster-level assignment and then to review each job in non-homogeneous clusters would require ~2000 decisions per expert, in contrast to evaluating 4255 unique questionnaire patterns or 14 983 individual jobs.

Conclusions: This proof-of-concept shows that using cluster models as a data reduction step to identify jobs with similar response patterns prior to obtaining expert ratings has the potential to aid rule-based assessment by systematically reducing the number of exposure decisions needed. While promising, additional research is needed to quantify the actual reduction in exposure decisions and the resulting homogeneity of exposure estimates within clusters for an exposure assessment effort that obtains cluster-level expert assessments as part of the assessment process.

KEYWORDS: case–control studies; diesel exhaust; hierarchical clusters; occupational exposures

INTRODUCTION

In case–control and population-based cohort studies, occupational information is typically collected using questionnaires that elicit the subjects' lifetime work histories and, in some studies, using job- and industry-specific modules to obtain detailed questions on work tasks, location, and the chemicals, tools, and equipment used. The responses are usually reviewed job-by-job by an exposure assessor to assign exposure estimates for the agents of interest. Because the information is collected systematically, expert-based decision rules are being developed that link questionnaire response patterns to exposure decisions to automate the exposure assessment (Fritschi *et al.*, 2009; Behrens *et al.*, 2012; Pronk *et al.*, 2012; Friesen *et al.*, 2013; Wheeler *et al.*, 2013; Carey *et al.*, 2014; Friesen *et al.*, 2014b; Peters *et al.*, 2014). These decision rules make the exposure assessment process more efficient and transparent. They also provide a mechanism to replicate decisions in other studies.

One of the challenges in developing decision rules is that there can be nearly as many permutations of questionnaire response patterns as there are jobs in the study. Many of the permutations are rare or not observed in a given study population, in part because skip patterns in the questionnaires result in only subjects with a particular response to a question (parent question) being asked nested follow-up questions (child questions). Thus, it is not efficient to develop rules for all potential patterns. In efforts to develop decision rules, exposure assessors generally use their judgment to identify a questionnaire response pattern that would suggest the same exposure estimates and, correspondingly, identify response patterns that would result in different exposure estimates. However, two or more exposure assessors independently developing decision rules may identify different questionnaire response patterns requiring exposure decisions because each may weigh the aforementioned information differently. As a result, a systematic, transparent data reduction step is needed.

One potential approach to reduce the number of questionnaire permutations that need evaluation, prior to employing expert opinion, is to use clustering methods to identify response patterns that are very similar, but not necessarily completely identical. Clustering methods can be used to identify natural groupings, or clusters, of observations when no outcome (i.e. an exposure metric) is available. A variety of statistical approaches to identify clusters of observations have been developed and applied in a wide range of settings, including genomics, identification of disease subgroups, handwriting recognition, and image segmentation (Garcia-Aymerich *et al.*, 2011; Maulik and Sarkar, 2013; Yoshioka *et al.*, 2013). The goal of clustering methods is to partition the observations into mutually exclusive clusters, such that observations that are similar are grouped into the same cluster and observations that are dissimilar are grouped into different clusters. In the only occupational example we found, Hines *et al.* (1995) used hierarchical clustering methods to identify that workers in a cohort of semiconductor workers with potential exposure to 14 correlated exposure agents could be organized into three distinct groups, where workers within a group were exposed to the same group of agents.

In this paper, we describe a proof-of-concept study that applies cluster models to occupational information available within a case–control study to group the nearly 15 000 jobs reported by study subjects into groups with similar questionnaire responses (i.e. cluster). We then evaluated whether the jobs within a cluster had similar exposure estimates using previously made exposure estimates (Pronk *et al.*, 2012). These evaluations were conducted as a first step in determining the utility of clustering models to facilitate rule-based occupational exposure assessment in case–control studies. We believe that the cluster models could be used, prior to obtaining expert estimates of exposure, to identify groups or 'clusters' within similar response patterns, as we report in a recent conference abstract (Friesen *et al.*, 2014a). The exposure assessment could then be conducted in two stages, where first the experts provide cluster-level estimates based on each cluster's profile of occupational responses, and then followed with a one-by-one expert review of a subset of jobs identified to be in a cluster that had variable exposure estimates, thereby potentially reducing the number of exposure decisions to be made. Because the utility of clustering methods in this context is unknown, our goal in this paper was to provide initial insights into the potential use of clustering methods to assist in the process of developing exposure decision rules for case–control studies and to provide support for using clustering methods as a component of future exposure assessment efforts. The application of a hierarchical cluster model to assist the exposure assessment process, including having experts provide cluster-level exposure assignments, will be reported separately.

MATERIALS AND METHODS

Model overview

Many clustering methods exist to group observations in the absence of an outcome measure (Hartigan, 1975; Hastie *et al.*, 2003; Everitt *et al.*, 2011). We chose agglomerative (bottom-up) hierarchical methods because they can easily incorporate the hierarchical nature of the responses to the occupational questionnaires (i.e. parent/child questions). In this bottom-up approach, each observation (i.e. job) begins as a separate 'cluster' (i.e. as many clusters as jobs). At each step, the clustering algorithm combines the two most similar clusters, continuing until all observations belong to a single cluster.

Before implementing a hierarchical clustering model, one chooses both a linkage criterion to determine which two clusters to combine at each step and a distance measure to evaluate the similarity of the input variables (here, questionnaire responses) between each observation (here, jobs). Several linkage methods are available in standard software, such as complete-linkage, average-linkage, and Ward's linkage methods (StataCorp L, 2009). We explored these three methods in preliminary analyses. Both complete-linkage and average-linkage methods identified only groups of jobs with exactly the same response and thus was nearly equivalent to evaluating all unique questionnaire response patterns; as a result, neither method was explored further. Ward's linkage was the only method of the three examined that led to dimension reduction, our main goal, and thus the only one examined in detail here.

Ward's linkage is most commonly used with the Euclidean squared distance measure (StataCorp L, 2009). At each step, Ward's linkage combines the two

clusters, C_m and C_k, into one cluster C_l that minimizes the total within-cluster squared error. The within-cluster quared error $S(C_l)^2$ of the lth cluster is defined as:

$$S(C_l)^2 = \sum_{i \in C_l} \sum_{p=1}^{P} \left(x_{ip} - \frac{1}{|C_l|} \sum_{j \in C_l} x_{jp} \right)^2$$

where $|C_l|$ is the number of observations in cluster C_l, P is the number of variables measured for each observation, x_{ip} and x_{jp} denote the pth variable corresponding to observations i and j. The total within-cluster squared error is the sum of within-cluster squared error over all K clusters (Ward, 1963; StataCorp L, 2009):

$$\sum_{l=1}^{K} S(C_l)^2$$

Study population and occupational exposure information

Our study included the 1213 cases and 1418 controls, who reported 14 983 jobs, from the New England Bladder Cancer Study (Colt *et al.*, 2011). Each subject completed a lifetime occupational history questionnaire that, for each job held, asked the job title, name and location of employer, type of service or product provided, year job started and stopped, work frequency (days per week, hours per day, months per year), principal work tasks and duties, tools and equipment used, and chemicals and materials handled. Two supplementary questions were also asked for each job: 'while on this job, 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?' For 64% of the jobs, information provided in the occupational histories triggered any of 67 job- or industry-specific modules that asked the subject more detailed questions about tasks and work activities related to exposure to diesel exhaust and other agents.

The occupational history and module questions were previously reviewed to identify questions that were directly or indirectly related to diesel exhaust exposure as part of the process to develop and extract decision rules, and is described elsewhere (Friesen *et al.*, 2013; Wheeler *et al.*, 2013). Briefly, the occupational histories provided three types of exposure information: (i) subject self-report of whether he or she worked near diesel or other engines or smelled diesel or other engine exhaust in that job (one variable); (ii) job-related questions extracted from job, industry, and the open-ended questions, that resulted in variables such as 'job had traffic exposure' and 'job used diesel-fueled equipment' (50 variables); and (iii) standardized industry classification codes (SIC) and standardized occupational classification codes (SOC), which were dichotomized and restricted to those expected by an industrial hygienist to have >5% of workers exposed to diesel exhaust (at three-digit level: 87 diesel-exposed SIC variables; 71 diesel-related SOC variables, 1 variable indicating any of the 87 diesel-exposed SIC, 1 variable indicating any of the 71 SOC). The modules provided two types of exposure information: (i) assigned module [67 dichotomous variables indicating module, 1 variable indicating module included questions related to diesel exhaust (hereafter, diesel-relevant module)]; and (ii) diesel-relevant module questions (154 dichotomous or continuous variables).

Each variable does not necessarily represent a single question, as the same question could have been asked across multiple modules and were thus merged. Similarly, questions with categorical responses were converted into a set of dichotomous variables. Questions that were not asked of all subjects were usually coded into two variables, one that indicated whether the question was asked (any response = 1; not asked = 0) and a second that indicated the response to that question (i.e. 1 = yes or number >0, 0 = no, 0, or not asked). For the 50 questions that occurred across multiple modules, 'not asked' was assigned '−1' rather than coded as a second variable to maximize the discrimination between those asked or not asked these key questions.

We re-coded continuous variables to approximate a 0–1 scale (or −1 to 1 scale) so that all variables had approximately equal weights because cluster models can be sensitive to the scale of the variables (Hennig and Liao, 2013). Units of hours per week were divided by 40 to re-code the variable into units of 'proportion of 40 hour work week'; a small number of responses exceeded 1 when the subject reported working >40 h per week. Ordinal variables (i.e. low, medium, and high) were centered on the middle category and re-coded to −1, 0, and 1, respectively.

Cluster model development

Due to known structures within the data, the jobs were split into three subsets that varied in the amount and type of occupational information: (i) jobs without a diesel-relevant module that had no variable indicating a possible diesel exhaust scenario ($n = 5929$, 40%); (ii) jobs without a diesel-relevant module but with at least one variable with a response indicating a possible diesel exhaust scenario (hereafter, shortened to 'jobs without diesel-relevant modules', $n = 5396$, 36%); and (iii) jobs with a diesel-relevant module ($n = 3658$, 24%). Jobs in the first subset were assumed unexposed because of the absence of any variable suggesting an exposure scenario and therefore no clustering analysis was performed. Cluster models were developed separately for the latter two subsets.

Hierarchical cluster models were developed using the 'cluster wardslinkage' command with the Euclidean squared distance measure (L2) in Stata S.E. v11.2 (StataCorp LP, College Station, TX, USA) using the available occupational information, but no exposure estimate. Cluster models for jobs without diesel-relevant modules used all diesel-related variables obtained from the occupational history responses. Cluster models for jobs with diesel-relevant modules incorporated both the occupational history and module diesel-related variables.

Evaluating within-cluster variability in exposure

The greatest challenge when working with clustering methods is determining how many clusters is 'enough' because of the absence of an outcome variable. Although there are some stopping rules based on statistical properties (Tibshirani *et al.*, 2001), it is generally up to the user to identify a sufficient number of clusters depending on the purpose of analysis. We focused our stopping rules on a practical question: how many clusters would an exposure assessor be able to evaluate in a reasonable time period? Thus, in the evaluations described below, we varied where we truncated each cluster tree, such that the jobs in each subset were grouped into varying numbers of clusters ranging from 100 clusters representing broad groupings higher up in the cluster tree (at 10–15 min per scenario, estimated 2 days for cluster-level assessment) to 1000 clusters representing more detailed groupings lower in the cluster tree (estimated 6–8 weeks). These time estimates do not include the likely needed second stage review, one-by-one review of jobs in heterogeneous clusters, which will be a function of the number of clusters used and the ability of the cluster models to identify similar exposure scenarios based on questionnaire response patterns (described in the section 'Estimated number of exposure decisions').

To gain insight into whether the cluster models identified groups of jobs that would be assigned the same values for the exposure metrics, we used previously developed algorithm-based estimates of probability, intensity, and frequency of exposure (Pronk *et al.*, 2012) that were independent of the clustering model process. Briefly, Pronk *et al.* (2012) developed consensus decision rules for occupation-, industry-, source-, and time period-specific exposure scenarios and then linked those scenarios to questionnaire response patterns in the occupational histories and modules to obtain an algorithm-based probability, intensity, and frequency estimate for each job. Probability was estimated as the expected proportion of workers likely exposed to diesel exhaust for that scenario, with cut points of <5, 5–49, 50–79, and ≥80% and assigned values 0, 1, 2, and 3, respectively, in the calculations described below. Intensity was estimated as the average level of respirable elemental carbon (μg m^{-3}) after review of the published monitoring data. Frequency was estimated as the average number of hours per week exposed to diesel exhaust. The rules were provided as supplementary material in Pronk *et al.* (2012). These exposure estimates were not used in the cluster model building process.

We evaluated the within-cluster variability using the algorithm-based estimates in two ways. First, we dichotomized the probability metric based on two definitions of exposure status (≥5 versus <5% probability; ≥50 versus <50% probability) and defined a cluster as 'homogeneous' if ≥75% of the jobs were assigned the same exposure status. We then calculated the proportion of clusters that were homogeneous at varying numbers of clusters (100 to 1000). Because 'homogeneity' was an arbitrary construct, we also examined the impact of using more or less stringent definitions of homogeneity using cumulative distribution plots of the proportion of homogeneous clusters. Second, to assess the within-cluster variability for the metrics using a continuous scale, we calculated the intraclass correlation coefficient (ICC) between the job estimates and the cluster mean (mean of the job estimates for all jobs assigned the same cluster). ICC is

the ratio of the variability between cluster means to the total variability in the data, where the values approach 1 as clusters become more homogeneous. ICCs were calculated using one-way analysis of variance models developed separately for cluster sizes increasing in increments of 100 from 100 to 1000 clusters per subset, with the cluster ID used as the grouping variable. ICCs for the probability metric were calculated based on all jobs within the subset. ICCs for the intensity and frequency metric were calculated based on only the clusters with a mean probability rating (rounded to the nearest integer) of 2 or 3, reflecting medium or high probability of exposure. This restriction reflects a common practice of using the intensity and frequency metrics in epidemiologic analyses only when a strict definition of exposure is met (Kromhout and Vermeulen, 2001; Purdue *et al.*, 2011).

Estimated number of exposure decisions

As an approximation, we estimated the total number of potential exposure decisions that may be needed to assess diesel exhaust exposure assuming a two-stage assessment process. The number of exposure decisions was calculated as the sum of the number of clusters (representing the first stage, cluster-level expert review) and the number of jobs that were not in homogeneous clusters (representing second stage expert review, where jobs in heterogeneous clusters would likely need to be reviewed one-by-one).

RESULTS

The 5396 jobs without diesel-relevant modules had 201 extracted response variables that represented 1120 unique questionnaire response patterns of the diesel-relevant variables, with a median of 1 and a mean of 4.8 jobs per pattern. The 3658 jobs with diesel-relevant modules had 392 extracted response variables that represented 3135 unique questionnaire response patterns, with a median of 1 and mean of 1.2 jobs per pattern.

To illustrate the visual output and relationship between observations, in Fig. 1 we show the hierarchical linkage relationship (cluster tree) for each of the two subsets when the observations were grouped into 100 clusters (shown at the bottom row) and their relationship to other higher level clusters (moving from bottom to top). The vertical axis indicates the magnitude of the distance measure between adjacent clusters. The branch lengths decrease substantially as one moves a horizontal line down from the top of the tree, showing that 'neighboring' clusters are more dissimilar at the top levels of the tree and more similar as one moves downwards. The scale of the distance measures and the resulting shape of the trees differed by subset because of the differing number and type of variables included in each model. Analyses grouping the observation into >100 clusters go deeper into the tree than what can be shown visually.

The cluster models can be directly used to create profiles of the questionnaire response patterns for each cluster that could be used in the first stage of a rule-based, expert-based exposure assessment. To illustrate, Table 1 lists the mean values for selected variables for four clusters from a cluster tree that was truncated at 300 clusters for jobs without diesel-relevant modules. Values

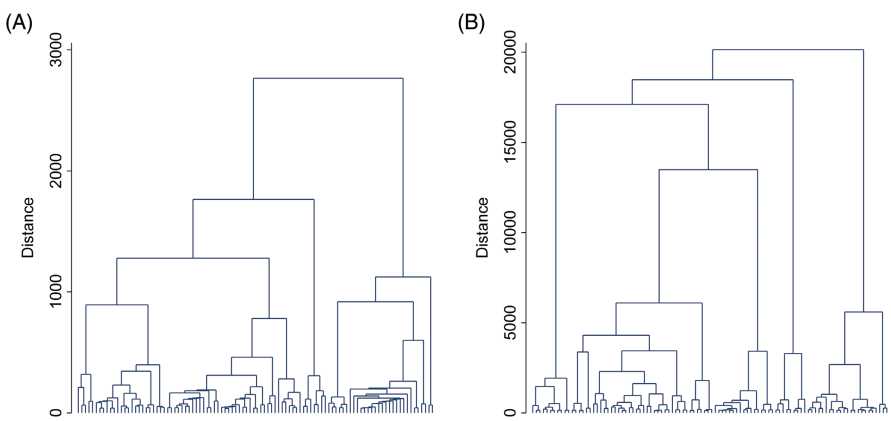

Figure 1 The top 100 clusters, with their Euclidean squared distance measures, for jobs without diesel-relevant modules (A) and with diesel-relevant modules (B).

Table 1. Cluster profiles with the mean response values for selected occupational history variables for four clusters identified from a cluster model for jobs without diesel-relevant modules, grouped into 300 clusters

Occupational history variable (subset)	Mean response value[a]			
	Cluster 7 ($n = 14$)	Cluster 236 ($n = 3$)	Cluster 237 ($n = 3$)	Cluster 238 ($n = 3$)
Free-text responses specifically said 'diesel'	0	0	0	0
Worked near engines or smelled engine exhaust	0.93	0	0	0.67
Traffic-related job, any[b]	0	1	1	1
Drivers	0	1	1	0
Firefighters	0	0	0	1
Parking attendants	0	0	0	0
Diesel-powered equipment, any[b]	1	1	1	1
Bus/truck	0	1	0	1
Heavy equipment	1	0	0	0
Industry with probable exposure, any[b]	1	0	0	0
Heavy construction	1	0	0	0
Logging	0	0	0	0
Industry with possible exposure, any[b]	0	0	0	0
Light construction	0	0	0	0
Military	0	0	0	0
Diesel-exposed SIC	1	0.33	1	0.67
Diesel-exposed SOC	1	0.67	1	1

[a]Mean response values: 1 = response for all jobs within cluster was 1 (yes); 0 = response for all jobs within cluster was 0 (no/not identified); >0 to <1 = response for that variable varied for jobs within cluster.
[b]'Any' refers to the presence of any category within this group. Specific categories were identified in separate variables.

of '0' indicate that all jobs within that cluster had a 'no' response for that variable; values of '1' indicate that all jobs had a 'yes' response. For instance, all jobs in cluster '237' had identical questionnaire response patterns for the selected variables, with a value of '1' (yes) to 'traffic-related job, any', 'drivers', 'diesel-powered equipment, any', 'diesel-exposed SIC', and 'diesel-exposed SOC' and a value of '0' (no/not identified) for the remaining variables listed in the table. Values >0 but <1 indicate varying amounts of heterogeneity in the responses to each variable within that cluster. For instance, the mean value for the variable 'worked near engines or smelled engine exhaust' for jobs in cluster '9' was 0.75, indicating that 75% of the jobs had a '1' (yes) response. If the heterogeneity occurred in a variable that an expert determined influenced the exposure decision, the assessor would need to move further down the cluster tree (i.e. increase the number of clusters) to separate disparate jobs or, alternatively, depending on the number of jobs and the amount of heterogeneity, evaluate each job within that cluster individually.

Within-cluster variability

Based on the dichotomized probability rating, the proportion of homogeneous clusters generally increased as the number of clusters used to group the jobs increased for both subsets, but the rate of improvement decreased after 400 clusters (Fig. 2).

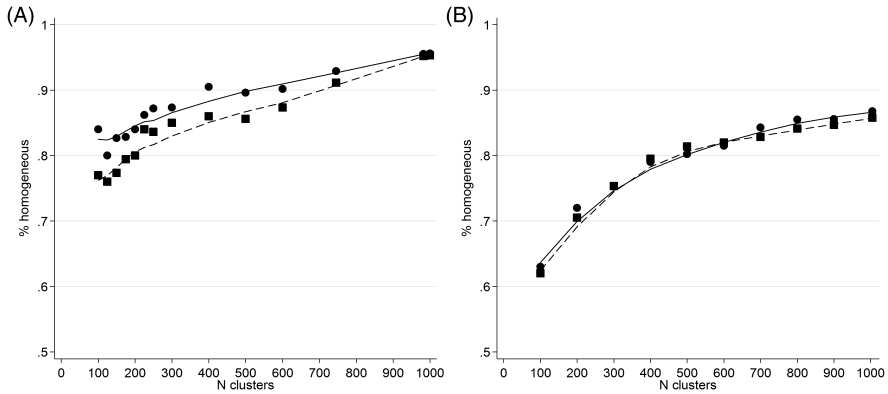

Figure 2 Proportion of homogeneous clusters (≥75% of jobs with same exposure status within cluster) with increasing number of clusters for jobs without diesel-relevant module (A) and for jobs with diesel-relevant modules (B). Dashed lowess (smoothing spline) lines = exposure status ≥5 versus <5% probability. Solid lowess lines = exposure status ≥50 versus <50% probability.

For jobs without diesel-relevant modules, 77–84% of the clusters were homogeneous when 100 clusters were used, increasing to 85–91% when 400 clusters were used. For this subset, the proportion of homogeneous clusters was somewhat higher when exposure status was defined as ≥50 versus <50% probability than for ≥5 versus <5% probability. For jobs with diesel-relevant modules, 62% were homogeneous when 100 clusters were used, increasing to 79–80% when 400 or 500 clusters were used. In this subset the definition of exposure status had little influence on cluster homogeneity (Supplementary Figure S1, available at *Annals of Occupational Hygiene* online). Cumulative distribution plots of the proportion of homogeneous clusters for the subset of jobs without diesel-relevant modules had a nearly flat slope between the values of 0 (all unexposed) and 1 (all exposed), indicating that the definition of homogeneity was robust (Supplementary Figure S1A, available at *Annals of Occupational Hygiene* online). The slope was steeper for the subset of jobs with diesel-relevant modules, indicating that the proportion homogeneous was more sensitive to the definition for this subset (Supplementary Figure S1B, available at *Annals of Occupational Hygiene* online).

For the ordinal probability and continuous intensity and frequency metric, ICCs reflecting the agreement between the job estimate and the cluster mean increased as more clusters were used to group the jobs (Fig. 3). The ICCs for all metrics and both subsets were generally >0.7 when 200 or more clusters were used, indicating very high agreement and minimal within-cluster variability, with one exception. For the frequency metric in the subset of jobs without diesel-relevant modules, ICCs ranged from 0.4 to 0.7 for 100–1000 clusters, suggesting more within-cluster variability for this metric and subset compared to the other metrics.

Estimated number of exposure decisions

The relationship between increasing the number of clusters and the expected number of exposure decisions that might be required, assuming a two-stage assessment, is non-linear (Fig. 4). The estimated number of decisions was minimized at ~800 decisions and 250 clusters for jobs without diesel-relevant modules (5% of jobs; 22% of the unique questionnaire response patterns) and at ~1150 decisions and 300 clusters for jobs with diesel-relevant modules (8% of jobs; 10% of unique questionnaire response patterns). Thus, we estimated that ~2000 exposure decisions (sum of number of decisions needed in both subsets) would be needed, rather than the 4255 decisions needed if every unique questionnaire response pattern was evaluated or the 14 983 decisions needed if each job was reviewed individually.

DISCUSSION

Hierarchical clustering methods were applied to diesel-relevant variables extracted from responses to

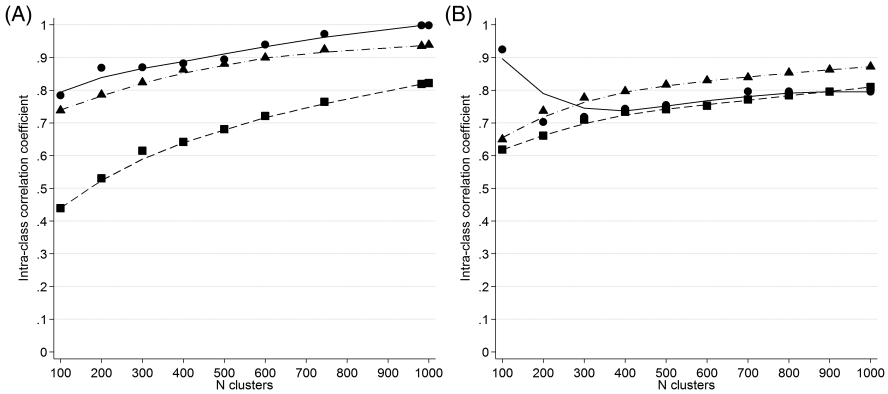

Figure 3 ICCs evaluating the agreement of the cluster-mean estimates to the algorithm-based estimates for probability (triangles), intensity (circles), and frequency (squares), for jobs without diesel-relevant module (A) and jobs with diesel-relevant modules (B). Analyses of the intensity and frequency estimates were restricted to clusters with a mean probability rating of medium or high.

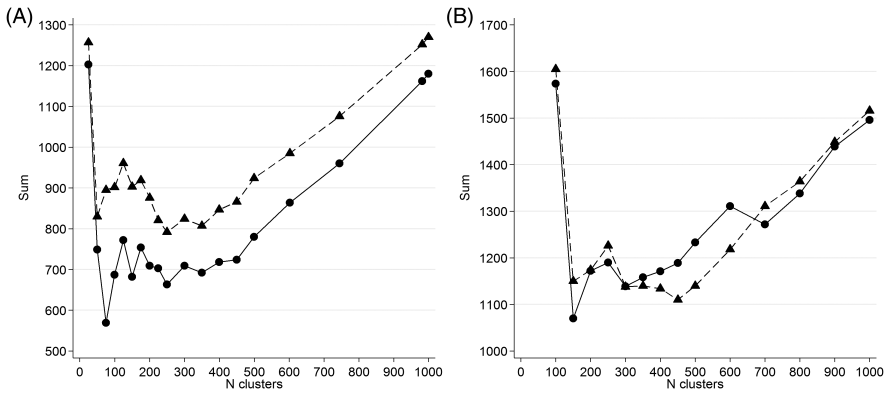

Figure 4 Sum of the number of clusters and the number of jobs in non-homogeneous clusters for jobs without diesel-relevant module (A) and jobs with diesel-relevant modules (B). Solid lines with circles = exposure status ≥5 versus <5% probability; dashed lines with triangles = ≥50 versus <50% probability.

occupational questionnaires to group jobs into clusters with similar questionnaire response patterns. Our evaluations comparing the resulting clusters to previously assigned probability, intensity, and frequency estimates suggest that the clusters, for the most part, identified groups of jobs with the same exposure status with minimal within-cluster variability (ICCs generally >0.7). Some heterogeneity in the clusters was expected, because some exposure scenarios occurred too rarely to fully distinguish into a separate cluster. In practice, we expect that heterogeneity within clusters could be accounted for in a two-stage exposure assessment approach that first assesses exposure at the exposure scenario level (cluster level), followed by review of individual jobs within expert-identified heterogeneous clusters (job level). We estimated that this two-stage approach can drastically reduce the number of exposure decisions required. Based on the insights gained in this proof-of-concept study, sufficient support was provided to move forward with using hierarchical cluster models as part of the exposure assessment and to conduct further evaluations into their use (i.e. Friesen *et al.*, 2014a).

Clusters became more homogeneous, with less within-cluster variability, as the number of clusters used increased. However, in our particular example, the sum of the number of clusters and the number of jobs in non-homogeneous clusters began rising more rapidly

when >300 clusters were used. Thus, simply increasing the number of clusters did not reduce the exposure assessment burden. Minimums of the sum occurred when the number of clusters used was ~5–8% of all jobs and 10–22% of the number of unique permutations of questionnaire responses. Not surprisingly, more clusters were needed for the jobs with modules to identify similar response patterns to create homogenous clusters than for jobs without modules, despite the fewer number of jobs represented by this subset, because of the greater amount of information available for those jobs (i.e. 154 additional variables).While it is possible to determine the optimal number of clusters when the exposure decision (i.e. outcome) is available, in practice the cluster models would be developed and used to help obtain the exposure decisions and thus the optimal number of clusters would not be known in advance. The approximations described above could be used to estimate the exposure assessment burden against resource constraints in future uses of cluster models. Simulation studies and additional sensitivity analyses (e.g. changing the distance measure or the linkage method) may also provide guidance.

A strength of this study was the preselection of the variables to those directly or indirectly related to the agent of interest (for instance, excluding SOC and SIC variables that were unrelated to diesel exhaust scenarios). Preliminary analyses that incorporated all available variables required more exposure decisions (data not shown), which is not surprising because of the greater amount of information included in the model and because the model is agnostic to whether the variable is relevant to the exposure of interest. Thus, the development of a hierarchical cluster model to identify similar questionnaire response patterns (exposure scenarios) is likely best suited to an evaluation of a single agent or a small number of agents where exposure decisions would be determined based on the same subsets of occupational questions. Consequently, we expect that it would be more efficient to develop separate cluster models for agents that rely on different subsets of questions.

Our evaluations of within-cluster variability were assessed based on algorithm-based exposure estimates rather than the estimates obtained from a one-by-one review of each job because our use of hierarchical cluster models was designed to closely replicate the algorithm-based assessment approach (Pronk *et al.*, 2012; Friesen *et al.*, 2013). Comparisons to a one-by-one review would likely identify more heterogeneity than was observed here because the one-by-one reviews consider the entire pattern of responses, not just the responses related to diesel exhaust exposure. However, comparisons to estimates from a single expert is not considered to be a gold standard and would not account for the natural variability between any two experts. Previous comparisons of the algorithm-based estimates in a subset of this dataset to estimates from a one-by-one review by three experts found very good agreement with the aggregate rating of the three raters (weighted kappa = 0.82) (Friesen *et al.*, 2013), providing strong support for using algorithm-based approaches to efficiently incorporate the estimates from multiple experts.

Our use of a hierarchical cluster model to identify diesel exhaust-related exposure scenarios provided important insights into differences in the models' performance by the type of information available (occupational history only or both occupational history and modules) and the impact of varying the number of clusters on the within-cluster variability and number of estimated exposure decisions. Our evaluations also had several limitations. First, the evaluations did not account for exposure assignment differences for jobs with similar response patterns but held in different time periods, and thus likely underestimates the proportion of homogeneous clusters if time period-specific cluster estimates were going to be obtained. In practice, we would ask experts to provide cluster-level assignments for multiple time periods predefined by major changes in exposure, regulations or other changes in use patterns (e.g. pre-1980, 1980–1994, 1995+) (e.g. Friesen *et al.*, 2014a). Second, our evaluations of the number of potential exposure decisions needed assumed that an exposure assessor would identify the same clusters as heterogeneous as those identified by the exposure status variables; however, this would vary based on the exposure assessor, the agent, and the extent of the variability among the variables for each cluster. Third, the number of decisions required may be overestimated because we may be able to capture some within-cluster heterogeneity by directly using the response to a variable in a programmable decision rule, rather than a one-by-one job review. For example, if the frequency estimate was directly related to the response to the question 'how many hours a week did you drive a truck', the response to that question for each subject could be directly linked to that subject's frequency estimate in automated decision rules. Lastly, the impact of variable coding on the identification of

clusters requires additional study. For instance, we chose not to normalize the variables (i.e. mean = '0') so that the responses remained interpretable for the experts and because the data was highly skewed toward 0 because '1's' (yes responses) were not prevalent so the mean overall response was near 0 for most variables. For the 50 diesel-relevant variables asked across multiple modules, the mean was <0 because '−1' was assigned when that question was 'not asked' (ranged from 49 to 99% not asked) rather than indicated in a separate variable because we wanted to maximize the discrimination between those asked or not asked these key questions. Additionally, one could consider weighting schemes that provided more weight for variables *a priori* expected to be highly predictive of exposure.

In summary, our evaluations provide a proof-of-concept that hierarchical clustering methods can systematically identify exposure scenarios represented by groups of jobs with similar questionnaire response patterns and similar exposure estimates when a large number of occupational questions are available electronically and coded systematically. Implementing this approach is expected to help us more efficiently use multiple exposure assessors to assess the same exposure scenarios. The implementation of this approach to help a team of exposure assessors to assess an exposure agent based on questionnaire responses will be evaluated in future studies.

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

The authors declare that there are no conflicts of interest relating to the material in relation to this paper.

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 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 JS, Karagas MR, 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.

Everitt BS, Landau S, Leese M *et al.* (2011) *Cluster analysis*. 5th edn. London, UK: John Wiley & Sons, Inc..

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.

Friesen MC, Locke SJ, Zaebst D *et al.* (2014a) Using machine learning to efficiently use multiple experts to assign occupational lead exposure estimates in a case-control study. *Occup Environ Med*; 71 (Suppl. 1): A25–6. doi:10.1136/oemed-2014-102362.79

Friesen MC, Park D, Colt JS et al. (2014b) Developing estimates of frequency and intensity of exposure to three types of metalworking fluids in a population-based case-control study of bladder cancer. *Am J Ind Med*; 57: 915–27.

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. doi:10.1155/2009/957023

Garcia-Aymerich J, Gómez FP, Benet M *et al.* (2011) Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax*; 66: 430–7.

Hartigan JA. (1975) *Clustering algorithms*. New York, NY: John Wiley & Sons, Inc.

Hastie T., Tibshirani R, Friedman J. (2003) *The Elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Hennig C, Liao TF. (2013) How to find an appropriate clustering for mixed type variables with application to socio-economic stratification. *Appl Statist*; 62 (Part 3): 1–23.

Hines CJ, Selvin S, Samuels SJ *et al.* (1995) Hierarchical cluster analysis for exposure assessment of workers in the Semiconductor Health Study. *Am J Ind Med*; 28: 713–22.

Kromhout H, Vermeulen R. (2001) Application of job-exposure matrices in studies of the general population: some clues to their performance. *Eur Respir Rev*; 11: 80–90.

Maulik U, Sarkar A. (2013) Searching remote homology with spectral clustering with symmetry in neighborhood cluster kernels. *PLoS One*; 8: e46468.

Peters S, Glass D, Milne E *et al.*; 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: 215–9.

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.

Purdue MP, Bakke B, Stewart P *et al.* (2011) A case-control study of occupational exposure to trichloroethylene and non-Hodgkin lymphoma. *Environ Health Perspect*; 119: 232–8.

StataCorp L. (2009) *Stata multivariate statistics reference manual. Release 11*. College Station, TX: Stata Press.

Tibshirani R, Walther G, Hastie T. (2001) Estimating the number of clusters in a data set via the gap statistic. *J Roy Stat Soc B*; 63: 411–23.

Ward JHJ. (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*; 58: 236–44.

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.

Yoshioka S, Tsukamoto Y, Hijiya N *et al.* (2013) Genomic profiling of oral squamous cell carcinoma by array-based comparative genomic hybridization. *PLoS One*; 8: e56165.