Annals of Work Exposures and Health, 2021, 682–693
doi: 10.1093/annweh/wxab002
Advance Access publication 23 April 2021
Original Article

BOHS The Chartered Society for Worker Health Protection
OXFORD

Original Article

Testing and Validating Semi-automated Approaches to the Occupational Exposure Assessment of Polycyclic Aromatic Hydrocarbons

Albeliz Santiago-Colón[1], Carissa M. Rocheleau[1,*], Stephen Bertke[1], Annette Christianson[1,2], Devon T. Collins[3,4], Emma Trester-Wilson[3], Wayne Sanderson[3], Martha A. Waters[1], Jennita Reefhuis[5] and the National Birth Defects Prevention Study

[1]Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, 1090 Tusculum Ave, Cincinnati, OH 45226, USA; [2]Department of Environmental and Public Health Sciences, University of Cincinnati, Kettering Lab Building, 0056 160 Panzeca Way, Cincinnati, OH 45267-0056, USA; [3]Department of Epidemiology, University of Kentucky, College of Public Health, 111 Washington Avenue, Lexington, KY 40536, USA; [4]Inova Fairfax Medical Campus, 8110 Gatehouse Road, Falls Church, VA 22042, USA; [5]Centers for Disease Control and Prevention, National Center on Birth Defects and Developmental Disabilities, 4770 Buford Highway, MS-S106-3, Atlanta, GA 30341, USA

*Author to whom correspondence should be addressed. e-mail: CRocheleau@cdc.gov

Submitted 26 February 2020; revised 12 November 2020; editorial decision 11 December 2020; revised version accepted 7 January 2021.

Abstract

Introduction: When it is not possible to capture direct measures of occupational exposure or conduct biomonitoring, retrospective exposure assessment methods are often used. Among the common retrospective assessment methods, assigning exposure estimates by multiple expert rater review of detailed job descriptions is typically the most valid, but also the most time-consuming and expensive. Development of screening protocols to prioritize a subset of jobs for expert rater review can reduce the exposure assessment cost and time requirement, but there is often little data with which to evaluate different screening approaches. We used existing job-by-job exposure assessment data (assigned by consensus between multiple expert raters) from a large, population-based study of women to create and test screening algorithms for polycyclic aromatic hydrocarbons (PAHs) that would be suitable for use in other population-based studies.

Methods: We evaluated three approaches to creating a screening algorithm: a machine-learning algorithm, a set of *a priori* decision rules created by experts based on features (such as keywords) found in the job description, and a hybrid algorithm incorporating both sets of criteria. All coded jobs held by mothers of infants participating in National Birth Defects Prevention Study (NBDPS) ($n = 35{,}424$) were used in developing or testing the screening algorithms. The job narrative fields considered for all approaches included job title, type of product made by the company, main activities

Published by Oxford University Press on behalf of The British Occupational Hygiene Society 2021.

What's important about this paper?

Advances in machine learning and predictive analytics offer promising opportunities to improve exposure assessment. This project explored how an open-text screening method based on a machine-learning algorithm compared to *a priori* screening rules developed by expert opinion and a hybrid method incorporating both approaches with respect to identifying jobs with potential exposure to polycyclic aromatic hydrocarbons (PAHs). The open-text screening method has the highest specificity, but the lowest sensitivity of the methods tested. The hybrid method performed similarly to the rules based on expert opinion, but could be optimized to improve sensitivity. The methods demonstrated can be applied to determine PAH exposures in other population-based studies with job narratives.

or duties, and chemicals or substances handled. Each screening approach was evaluated against the consensus rating of two or more expert raters.

Results: The machine-learning algorithm considered over 30,000 keywords and industry/occupation codes (separate and in combination). Overall, the hybrid method had a similar sensitivity (87.1%) as the expert decision rules (85.5%) but was higher than the machine-learning algorithm (67.7%). Specificity was best in the machine-learning algorithm (98.1%), compared to the expert decision rules (89.2%) and hybrid approach (89.1%). Using different probability cutoffs in the hybrid approach resulted in improvements in sensitivity (24–30%), without the loss of much specificity (7–18%).

Conclusion: Both expert decision rules and the machine-learning algorithm performed reasonably well in identifying the majority of jobs with potential exposure to PAHs. The hybrid screening approach demonstrated that by reviewing approximately 20% of the total jobs, it could identify 87% of all jobs exposed to PAHs; sensitivity could be further increased, albeit with a decrease in specificity, by adjusting the algorithm. The resulting screening algorithm could be applied to other population-based studies of women. The process of developing the algorithm also provides a useful illustration of the strengths and potential pitfalls of these approaches to developing exposure assessment algorithms.

Keywords: exposure assessment; female worker; jobs; machine-learning algorithm; National Birth Defects Prevention Study; occupation; prediction model; polycyclic aromatic hydrocarbons; population-based; regularized logistic regression

Introduction

Occupational exposure can be captured in various ways in epidemiological studies. Direct measurement of a participant's exposure, or biomonitoring of their absorbed dose, is considered the gold standard; however, these are not always possible to obtain (Rezagholi and Mathiassen, 2010). For studies enrolling participants from a wide geographic area or multiple workplaces, it may not be practical to obtain direct measurements. For most retrospective studies (e.g. retrospective cohort or case–control studies), it is unattainable to capture direct measurements or conduct biomonitoring. In these studies, occupational exposures can be estimated retrospectively via self-reported exposure, job-exposure matrices (JEMs), or review by one or more expert raters of detailed job descriptions to assign estimates of exposure (Clavel *et al.*, 1993; Fleming *et al.*, 2014; Fritschi, 2019). Each method has limitations.

Self-reported exposure (i.e. asking a participant if they were exposed to an agent) is the simplest method, but also tends to be the least accurate due to limits in the participant's knowledge about the chemical, as well as the time elapsed between the exposure and the interview, which can affect recall. Most participants struggle to report precise information on which chemicals or equipment they used, estimate amounts of exposure, or provide information on ventilation in their workplace. Consequently, this approach generally has low validity, which can result in incorrect characterizations of true exposure–effect associations (Teschke *et al.*, 2002). However, participants are generally able to accurately self-report a job description. A job description typically consists of their job title, employer, what the employer makes/does, and their typical job duties.

These job descriptions can be matched to industry and occupation codes, which can, in turn, be matched to exposure estimates via a JEM. Although JEMs are efficient, reproducible, and standardized (and often more valid than self-report), they do not account for variations in job tasks between individuals with the same job title; that is, all people with a given job title will be assumed to have the same exposure (Clavel *et al.*, 1993; Teschke *et al.*, 2002; White *et al.*, 2008; Fritschi *et al.*, 2009; Peters *et al.*, 2014; Fritschi, 2019). To reduce misclassification of exposure and account for a participant's individual-level variation in job tasks, expert raters can review each job description to assign likely exposure. Using multiple independent raters (typically experienced industrial hygienists familiar with job tasks and exposures) and resolving any disagreements by a consensus process helps increase the validity of this method. This approach can account for variations in job tasks and is largely considered the most valid for retrospective exposure assessment (Fritschi *et al.*, 2009). However, expert review of detailed job descriptions is costly, time-consuming, and the quality of the expert rater review can vary (Rocheleau *et al.*, 2011; Pronk *et al.*, 2012; Wheeler *et al.*, 2013; Florath *et al.*, 2019).

To reduce the cost and time required for exposure assessment in a large population with low exposure prevalence, various methods of screening or "triage" are often used to quickly reduce the number of jobs to be reviewed by an expert; job descriptions that do not meet certain screening criteria are assumed to be unexposed without the need for further manual review. These methods can reduce the burden of expert raters by identifying clearly unexposed participants, therefore raters only review a small cohort of job descriptions with greater potential for exposure. In addition, these methods could provide a more standardized method of conducting exposure assessment and could reduce the time of exposure assessments overall, while providing comparable quality (Wheeler *et al.*, 2013; Friesen et al., 2015, 2016; Wheeler *et al.*, 2015; Florath *et al.*, 2019; Sauve *et al.*, 2019) if carefully validated and applied.

In developing screening algorithms, some studies have used *a priori* expert decision rules based on industry and occupation (North American Industry Classification System, NAICS, and Standard Occupational Classification, SOC, codes), keywords, or both (Fritschi et al., 1996, 2009; Friesen et al., 2013, 2014, 2015, 2016; Fritschi, 2019). While decision rules increase efficiency, they are inherently subjective; when they are not well documented or validated, there is a potential for errors or over/underestimation of the likely accuracy of the exposure assessment. If existing exposure data are available, machine learning can use these data to derive algorithms for predicting exposure (i.e. extracting decision rules from the data) in an objective way. The major challenge of machine learning is having adequate data for the algorithm to learn from. Since the algorithm cannot be informed by data that is not in the training data set, a machine-learning algorithm may perform poorly when it encounters a new or unfamiliar situation. Machine learning can be incredibly efficient and develop highly successful predictive algorithms when it is informed adequately by data, however. Multiple groups have approached machine learning for the classification of narratives in public health data (Alpaydin; Lehto *et al.*, 2009; Bertke et al., 2012, 2016; Measure, 2014; Marucci-Wellman *et al.*, 2015; Vallmuur, 2015).

Recently, the occupational exposure assessment for polycyclic aromatic hydrocarbons (PAHs) was completed for the National Birth Defects Prevention Study (NBDPS); this assessment used a manual review of each job description by one or more trained raters. PAHs are a group of persistent chemicals formed during the incomplete combustion of organic substances (Agency for Toxic Substances and Disease Registry, 1995; U.S. Environmental Protection Agency, 2008) and can be found in a wide variety of settings, including the workplace.

This study explored various exposure assessment screening strategies using the NBDPS exposure assessment of PAHs from 1997 to 2011. Strategies included a machine-learning algorithm, expert *a priori* decision rules that included keywords and NAICS/SOC codes, and a combination of both. The results of these differing approaches were compared to each other and to the results of a manual expert rater assessment. The purpose of this manuscript is to (i) illustrate the application of different screening approaches (using either a machine-learning approach versus an expert-developed set of screening rules), and (ii) to develop and validate a screening algorithm for PAH exposure that is generally comparable to multiple expert rater review that may be used in future population-based studies of women, such as BD-STEPS (Tinker *et al.*, 2015).

Methods

Study population

The NBDPS is the largest population-based case–control study ever conducted in the United States (US) examining risk factors for over 30 structural birth defects. Full study details have been published elsewhere (Yoon *et al.*, 2001; Reefhuis *et al.*, 2015).

Participating mothers (n = 44,029) completed a computer-assisted telephone interview (CATI) in English or Spanish within 2 years of their index child's birth. The interview covered many demographic, maternal health, and lifestyle questions, including detailed descriptions of jobs held for 1 month or longer during the three months prior to conception until the end of the pregnancy (B3-P_{END}). This study considered all jobs (n = 35,424) held by mothers who were employed for one month or more during the study period (n = 30,456 women). For women completing the interview in Spanish, bilingual interviewers provided English translations of free-text responses (including the job description) in the study database.

The occupational section of the interview inquired about the employer or company's name, what the company made/did, the mother's job title, her typical job duties or tasks, any equipment or chemicals she used, the typical number of hours she worked at that job on a work day, the typical number of days worked per week, the month and year she started the job, and the month and year she stopped the job. Institutional review boards (IRB) at the Centers for Disease Control and Prevention and each participating site approved the study protocols, and all participants provided informed consent.

Exposure assessment

The case–control status of having offspring with birth defects was not considered in this analysis since it does not affect whether a given job might have had exposure to a given chemical and exposure. However, all exposure assessments were blinded to case–control status. Prior to the PAH-exposure assessment, qualified industry and occupation coders assigned North American Industry Classification System (NAICS) and Standard Occupational Codes with version information to each job description provided by a participating mother; codes were standardized to 2007 NAICS and 2010 SOC at the end of the study for the final database release. Final quality control of the NAICS and SOC coding was performed by a separate group of industrial hygienists.

During the initial PAH-exposure assessment for NBDPS, each job was reviewed by at least one trained screener to "screen out" jobs considered to have a high confidence of no exposure potential. Two industrial hygienists (expert raters) further evaluated the narrative job description for jobs flagged as possibly exposed to PAHs by the trained reviewers. A third industrial hygienist resolved by consensus any discrepancies in the assigned ratings between hygienists. Raters assigned each job to a PAH-exposure category (none, direct exposure only, indirect exposure only, and both direct plus indirect exposure). Only exposure to PAHs from occupational sources was considered; mothers' self-reported exposure to secondhand smoke at work and home are ascertained elsewhere in the interview. For this analysis, the expert rater assessment was considered the "gold standard," and the exposure was a dichotomous (yes/no) variable.

NBDPS enrolled participants on an ongoing basis for more than 15 years, therefore, exposure assessment was conducted at two different time points. The first exposure assessment, covering birth years 1997–2002, included a duplicate review of each job description by two industrial hygienists (expert raters) with discrepancies resolved by consensus. The second exposure assessment, covering birth years 2003–2011, used an initial screening step to increase efficiency. To ensure quality in the exposure assessment, all industrial hygienists extensively reviewed existing PAH-exposure literature, including health hazard evaluations and monitoring data. To improve consistency between the first and second exposure assessments, screeners and raters for the second exposure assessment rated two sets of 100 jobs from the first exposure assessment, with the review of discrepancies between each other and the original ratings. To assure comparability between the screeners, a 5% random sample was also selected for duplicate screening by all screeners. A third project manager also reviewed the 5% sample for quality control, and reviewed additional jobs flagged for quality control (e.g. those that were rated discordantly from 75% or more of similar SOC codes).

The final database of job descriptions and consensus PAH-exposure status (possibly exposed above baseline, unexposed above population baseline) generated from a job-by-job review by multiple expert raters was used to create a screening algorithm that could be applied to future population-based studies of women. In the process, we compared screening algorithms created using a machine-learning algorithm, manual decision rules, and a hybrid (combination) of the two approaches.

Screening approaches evaluated

Machine-learning algorithm screening

This study employed a type of regularized logistic regression (known as RIDGE regression) for the machine-learning approach. Details of this method are described elsewhere (Bertke *et al.*, 2016). In brief, this procedure uses logistic regression to calculate the probability of an event (e.g. exposure to PAHs) based on the features of a job narrative, but applies a regularization penalty parameter to avoid over-fitting the data. In this study, features were the occurrence of single words, occurrence of two-word sequences, and 3-digit industry (NAICS) and occupational (SOC) codes (individually and in combination). To be eligible as a feature, each word or sequence of words had to occur in at least three jobs in the training set. The job narrative fields considered for this algorithm included job title, type of product made by the company, main activities or duties, and chemicals or substances handled. Fuzzy matching for misspellings (i.e. diesel, deisel, and desel are considered different words) and run-on words was not performed. Furthermore, singular versus plural words and verb tenses were categorized as different features. All words were converted to lowercase before running the algorithm to avoid having a word considered as two different features due to differences in word case. Common, less informative words (e.g. "the," "my," "a," "an," among others) were not treated as features.

All jobs were randomly split into a training set (n = 30,424 jobs) and a testing set, also referred to as a validation set (n = 5000 jobs). The prediction model was built from the training set by considering the occurrence of features (keywords and industry/occupation codes) in the job description narrative and running the logistic model on these data. Every feature was assigned a weight or beta regression coefficient. The weights of the features present in each job description were summed and using this information, the probability of exposure was calculated. Overall, large positive weights represent features often associated with exposure and large negative weights represent features often associated with an unexposed job. From this process, we were able to see which features had large positive weights and if they were associated with an exposed job. Afterwards, the prediction model was applied to the testing set. The RIDGE used the relationship between the exposure category and the different features in the training set to return a predicted probability (0–1) of exposure for each job in the testing set. By default, jobs with an assigned probability of 0.50 or higher were considered exposed and with a calculated probability of <0.50 were considered unexposed. Those considered as exposed could be then assigned for manual review. The prediction model was developed and implemented with Python version 3.1 (Copyright © 2001–2017 Python Software Foundation; All rights reserved https://www.python.org).

Comparisons of the RIDGE results were performed in quintile and decile distributions of the probability of exposure calculated by the RIDGE with the results of the expert rater assessment at each level. Based on which probability of exposure level returned the most discrepancies (i.e. false negatives) with the expert rater assessment, the probability of exposure levels that could serve as the new probability of exposure cutoffs were identified. To illustrate the effect of choosing different probability score thresholds for manual review, two probabilities of exposure cutoffs were selected using the quintile and decile distributions of the probability of exposure for jobs classified as exposed and unexposed by the RIDGE. For quintile and decile distributions, the top level of jobs classified by the RIDGE as unexposed had the largest number of false negatives. As a result, instead of the RIDGE assigning a job for manual review when the probability of exposure was 0.50 or higher, the new cutoffs represented a new probability level for a job to be referred for manual review. All evaluations were conducted in SAS version 9.4 (Copyright (c) 2002–2012 by SAS Institute Inc., Cary, NC, USA).

a priori screening rules procedure

This approach used a strategy commonly used to streamline exposure assessments in which no prior data are available: to work with experts to develop a list of keywords, combinations of words, and/or job codes (e.g. NAICS and SOC codes or combinations) that could indicate possible exposure. If the job description flagged with either a keyword or NAICS/SOC that indicated potential exposure, the job description was assigned for manual review. Jobs not flagged by keyword or NAICS/SOC were assumed to be unexposed under this approach. We used two certified industrial hygienists, each with >20 years of experience in exposure assessment (MAW, WTS) but who had not conducted the job-by-job exposure coding for NBDPS (though both were familiar with the data set and ratings) to identify 84 features: 73 keyword rules and 11 industry codes. Keywords included root words (e.g. cook-, -fry-, weld-, grill-), combinations of words (e.g. coal tar, iron making), and single words (e.g. burning, gasoline, exhaust). Industry codes ranged from 2-digit to 5-digit codes. The full list of features is available upon request.

Hybrid procedure

Figure 2 and Supplemental Figures 1–2 (available at *Annals of Occupational Hygiene* online) depict the hybrid approach. It was a combination of the machine-learning algorithm and the *a priori* screening procedures. First, the RIDGE approach is applied using the same default probability of exposure levels used by the RIDGE approach only (0.50), as well as the two additional cutoffs tested by this approach (0.10 or 0.04). Jobs flagged by the RIDGE as exposed, would then be assigned for manual review. For the jobs considered unexposed under the RIDGE approach, the *a priori* screening approach was applied. If features were flagged by the *a priori* screening, the flagged jobs would be assigned for manual review. Jobs not flagged by this second approach would be assumed as unexposed.

Evaluation of exposure screening procedures

The manually coded jobs in the testing set were used as a reference, with the assumption that the consensus ratings by the expert coders were correct. Each approach was compared to the manually coded jobs in the testing set and against each other. For each screening method and the hybrid of both screening methods, the authors summarized the number of jobs flagged for review and the number assumed unexposed, and how many of each category were considered exposed and unexposed compared to the gold standard of manual review of each job. Sensitivity, specificity, positive predictive value (PPV), and negative predictive values (NPV) were calculated for each approach, assuming the manual review of each job represented the true exposure of the job. This allowed the evaluation of trade-offs in terms of efficiency (i.e. fewer jobs requiring manual review) and quality (i.e. catching jobs for manual review that were considered to have exposure).

Results

RIDGE procedure and RIDGE cutoffs

For this study, the RIDGE considered over 30,000 features in the training set. These features contained 6494 single words (taken from the job description fields about job title, job duties, what the company makes/does, chemicals/substances or machinery handled); 118 NAICS codes (3-digit); 106 SOC codes (3-digit); and 2361 NAICS/SOC combinations.

From the 5000 jobs on the validation set, two jobs were excluded because the expert rater classification for those jobs was "unable to code" (jobs for which there was not enough information to assign a PAH-exposure rating); therefore, results from 4998 were evaluated on the validation set. Table 1 describes the agreement between the RIDGE screening algorithm and the gold standard of manual expert rater review. The RIDGE assigned most unexposed jobs a probability of exposure of 0.10 or less, while most exposed jobs had a probability of exposure of 0.80 or higher. Figure 1 describes how sensitivity changes the percentage of jobs flagged for manual review. As the number of jobs manually coded increases, so does the sensitivity.

Among jobs rated as unexposed (using the RIDGE probability cutoff ≤50%), the nearest decile of unexposed (probability 0.10–0.50) had 31.1% misclassified compared to the gold standard (Table 2). While in the next decile (0.05–0.10), 6.4% of the jobs were misclassified as unexposed. This illustrates that flagging probabilities near the cutoff for quality control review has the greatest efficiency in identifying false negatives. Based on these findings, a moderate and a conservative cutoff were created for the RIDGE to classify jobs as exposed. The moderate cutoff used a probability of exposure of 0.10 or higher to flag a job as potentially exposed, and the conservative cutoff used the probability of exposure of 0.04 or higher.

The RIDGE default cutoff flagged 467 (9.3%) jobs as potentially exposed, of which 381 (67.5%) jobs were correctly identified as exposed. Both the RIDGE moderate and conservative cutoffs performed better than the RIDGE default cutoff in correctly flagging exposed jobs, 92.4 and 98.2%, respectively. The RIDGE conservative cutoff was the most sensitive (98.2%), while the RIDGE high cutoff was the most specific (98.1%).

Table 1. Summary of the performance of the machine-learning algorithm, the keywords approach, and both methods combined

Comparison*	Expert-assigned code	RIDGE, *N* (%)	Features, *N* (%)	Hybrid 1,[a] *N* (%)	Hybrid 2,[b] *N* (%)	Hybrid 3,[c] *N* (%)
True positive	564 (11.3%)	381 (7.6)	482 (9.6)	491 (9.8)	537 (10.7)	556 (11.1)
True negative	4434 (88.7%)	4348 (87.0)	3956 (79.1)	3951 (79.1)	3790 (75.8)	3360 (67.2)
False positive	—	86 (1.7)	478 (9.6)	483 (9.7)	644 (12.9)	1074 (21.5)
False negative	—	183 (3.7)	82 (1.6)	73 (1.5)	27 (0.5)	8 (0.2)
Sensitivity	—	67.7	85.5	87.1	95.2	98.6
Specificity	—	98.1	89.2	89.1	85.5	75.8
PPV	—	81.6	50.2	50.4	45.5	34.1
NPV	—	96.0	98.0	98.2	99.3	99.8

RIDGE = regularized logistic regression used for the machine-learning algorithm approach.

*Results are based on the assumption that the expert-assigned code is correct.

[a]Hybrid 1 = RIDGE*Keyword. Cutoff when RIDGE probability of exposure ≥ 0.50 (default).

[b]Hybrid 2 = RIDGE*Keyword. Cutoff when RIDGE probability of exposure ≥ 0.10 (moderate).

[c]Hybrid 3 = RIDGE*Keyword. Cutoff when RIDGE probability of exposure ≥ 0.04 (conservative).

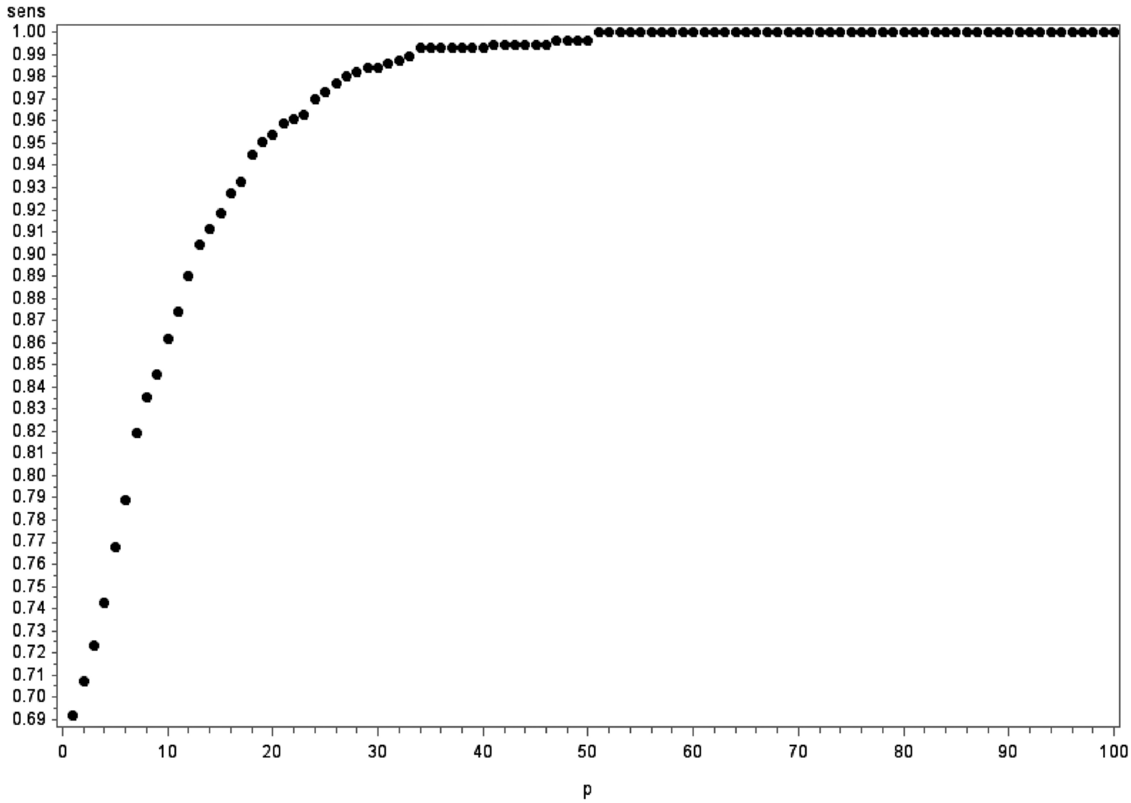

Figure 1. Percentage of job narratives flagged for review (p) versus sensitivity (sens).

a priori decision rules procedure

This procedure flagged 960 jobs (19.2% of all jobs), of which 482 (85.5% of all exposed jobs as classified by the expert rater) were correctly flagged as exposed. Sensitivity and specificity were 85.5 and 89.2%, respectively, for this approach (Table 1). The evaluation of exposed jobs missed by this procedure allowed the identification of potential words, combination of words, and occupational codes that should be considered when evaluating PAH exposures on job descriptions. Some features (words such as fryer and diesel) overlapped with keywords in the *a priori* approach.

Hybrid procedure

Figure 2 and Supplemental Figures 1–2 (available at *Annals of Occupational Hygiene* online) demonstrate the hybrid procedure at three probabilities of exposure cutoffs (0.50, 0.10, 0.04). Initially, job descriptions were evaluated by the RIDGE, and those with a probability of exposure of 0.50 or higher were flagged (n = 468, 9.4%), which assigned them to be manually reviewed (Fig. 2). The rest of the jobs' descriptions (n = 4530, 90.6%) were checked for features (keywords and NAICS/SOC codes) selected *a priori*. Jobs flagged by this process were also assigned to be manually reviewed (n = 506, 11.2% of all unflagged jobs). In summary, by reviewing 19.5% (n = 974) of all jobs in the validation set (n = 4998), this hybrid procedure was able to capture 87.1% (n = 491) of all exposed jobs

In comparison, the hybrid procedure with the RIDGE moderate cutoff (0.10) identified 95.2% (n = 537) of all exposed jobs by reviewing 23.6% (n = 1181) of all jobs in the validation set (Supplemental Figure 1, available at *Annals of Occupational Hygiene* online). The hybrid procedure using the RIDGE conservative cutoff (0.04) captured 98.6% (n = 556) of all exposed jobs by reviewing 32.6% (n = 1630) of all jobs in the validation set (Supplemental Figure 2, available at *Annals of Occupational Hygiene* online).

Discussion

Using a combination of known approaches for retrospective exposure assessment, this study was able to develop and test a prediction model using a machine-learning algorithm approach in a general population of

Table 2. Decile distribution of jobs classified as exposed or unexposed by the machine-learning algorithm: summary of false negatives and false positives

RIDGE classification	Probability exposed	*N*	Incorrect classification *N* (%)
Unexposed			
False negative	0.00–0.00	453	0
	0.00–0.01	453	0
	0.01–0.01	453	0
	0.01–0.01	453	0
	0.01–0.02	453	0
	0.02–0.02	454	3 (0.7)
	0.02–0.03	453	1 (0.2)
	0.03–0.04	453	9 (2.0)
	0.05–0.10	453	29 (6.4)
	0.10–0.50	453	140 (31.1)
Exposed			
False positive	0.50–0.56	46	19 (41.3)
	0.56–0.60	47	15 (31.9)
	0.61–0.68	47	12 (25.5)
	0.68–0.74	47	10 (21.3)
	0.74–0.80	47	12 (25.5)
	0.80–0.85	47	6 (12.8)
	0.85–0.90	47	4 (8.5)
	0.90–0.94	47	5 (10.6)
	0.94–0.97	47	3 (6.4)
	0.97–1.00	46	0

employed mothers. With having the exposure status of all jobs, as assigned by expert raters, the authors were able to contrast exposure scores from expert raters with those assigned by the RIDGE, *a priori* decision rules, and hybrid procedure. Additionally, it was demonstrated how different thresholds of the probability of exposure for assigning jobs to be manually reviewed can affect the performance of the RIDGE. It was possible to quantify the number of jobs to be manually reviewed to capture the majority of exposed jobs correctly for the different exposure cutoff point.

Overall, the RIDGE correctly identified a large number of exposed jobs by flagging a relatively small percentage of the total jobs. Most false positives and false negatives were concentrated around the probability of exposure cutoffs of the RIDGE. Most of the exposed jobs missed by the RIDGE (i.e. false negatives) were found at the top quantile and decile distributions of the probability of exposure for jobs classified as unexposed by the RIDGE. Modifying the probability of exposure cutpoint allowed the RIDGE to capture correctly 25–30% more exposed jobs, and decrease the number of false negatives and false positives. The combination of the RIDGE with the *a priori* decision rules procedure performed better than any approach individually. Furthermore, the hybrid procedure with the RIDGE conservative cutoff (probability of exposure ≥0.04) had the highest sensitivity (98.6%) of all procedures or combination of procedures, but specificity declined to 75.8%.

The size of the training set depends on the type of data available and the purpose for which the prediction model is being created. Previous studies using a machine-learning algorithm to auto-code injury causation have used various training set sizes depending on the amount of data available and the prevalence on the injury of interest in that data (Lehto *et al.*, 2009; Bertke *et al.*, 2012; Marucci-Wellman *et al.*, 2015; Bertke *et al.*, 2016). Applying an automated method to a general population, derived from a highly selected population, can create biased exposure assignments and classification errors—particularly if researchers are interested in finding out whether adverse effects occur at levels well below current legal limits, or for chemicals that do not have legal limits defined. In our study, using a large population of employed women allowed for the development of a more comprehensive automated method for the exposure assessment of PAHs. Since the NBDPS population has a wide variety of jobs, the large size of the training set allowed the RIDGE to learn from many different types of jobs and as a result, improved the performance when applied to the testing set by reducing the noise from the job description narrative (Lehto *et al.*, 2009).

The workplace can be a major source of a wide variety of PAHs, and exposure can fluctuate by industry and occupation. The Occupational Safety and Health Administration (OSHA) airborne permissible exposure limit (PEL) for coal tar pitch volatiles, a type of PAH, is 0.2 mg/m^3 in an 8-h work shift (Agency for Toxic Substances and Disease Registry, 1995; Occupational Safety and Health Administration (OSHA), 2019). OSHA's PEL for PAH comes from data obtained from OSHA's compliance monitoring; thus, it is only performed in occupations where there is *a priori* suspicion of high exposure that might exceed PELs. JEMs and automated methods based on data from compliance monitoring are most likely predominated by data on coal tar pitch volatiles (Lee *et al.*, 2015), which can negatively affect the performance of these methods for other types of PAHs unless an effort has been made to incorporate expert decision rules that explicitly address this limitation. Both the expert rater approach to assess PAH exposure in the NBDPS and the automated methods used

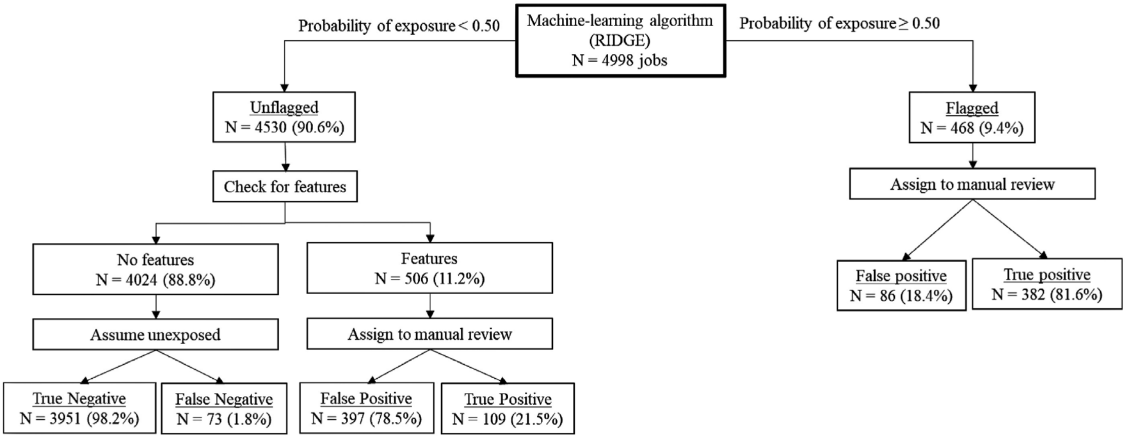

Figure 2. Hybrid procedure using the default probability of exposure cutoff (0.50). RIDGE refers to the regularized logistic regression; features are defined as words (single or two-word sequences), and 3-digit industry (North American Industry Classification System or NAICS) and occupational (Standard Occupational Classification or SOC) codes (individually and in combination).

in this study considered any PAH exposure above background levels for the classification of job exposure.

Future studies seeking to use this prediction model in a population highly exposed to PAHs or on a specific type of PAH can modify it to meet their exposure criteria. Modifications (e.g. adding keywords or industry codes) can also be made to update relevant job characteristics to account for changes in the workplace and job duties related to PAH exposure. Researchers seeking to evaluate a different type of exposure in a large population dataset can follow the methodology used for the development and application of this prediction model to create their own prediction model for any other exposure. Several considerations should be taken into account when developing a machine-learning algorithm based on resources available and the expectations for the prediction model (e.g. reduce the number of jobs that need manual review by identifying unexposed jobs with high confidence based on an investigator-identified threshold, such as specifying the RIDGE exposure probability cutpoint). Some of these considerations include how many resources will be available to manually review potentially exposed jobs; the cutoff probability level of exposure by the algorithm; and if the focus will be on occupation in general or on a specific type of occupation. Moreover, researchers using this prediction model in the future should consider manually checking a random sample of jobs near the cutoff point between exposed and unexposed classifications. It can help determine the need of modifying the probability of exposure cutoff point to capture more potentially exposed jobs. This study demonstrated how modifying the probability of exposure levels of the RIDGE can allow for a higher capture of exposed jobs by reviewing a small proportion of the data.

Although the development and improvement of machine-learning approaches has been ongoing for several years (Fritschi *et al.*, 2009; Bertke et al., 2012, 2016; Friesen et al., 2014, 2015, 2016; Vallmuur, 2015), its application on any particular dataset takes only minutes. Human raters must receive extensive training and over time, retraining is required to maintain consistency as the characteristics of the exposure and occupations evolve; this process takes monetary resources and time. When there are thousands of job narratives to evaluate, these coders/raters may be subject to fatigue (Marucci-Wellman *et al.*, 2015). The utilization of automated methods for retrospective exposure assessments helps to overcome many of these challenges, as fewer resources are needed for their development and implementation. Results showed that compared to expert rater assessment, both the RIDGE and the *a priori* decision rules performed well. Previous studies have shown some auto-coding approaches can achieve high levels of specificity overall (Williamson *et al.*, 2001; Lehto *et al.*, 2009), but sensitivity can vary. The exposure data used for this study had many more unexposed (89%) than exposed jobs; thus, it is expected for the RIDGE to learn more about unexposed jobs than from those exposed. Using a hybrid procedure, including *a priori* decision rules can complement the process and improve sensitivity by filling potential gaps occurring in the RIDGE. Results from the combination of both approaches showed that the inclusion of *a priori* decision rules enhanced sensitivity.

This study is not without limitations. The major limitation is that measurement data—a true gold standard for exposure—was not available; thus, any misclassification in the job-by-job coding by multiple raters is unknown and perpetuated in the models. Very few direct measurements have been published in occupations with PAH levels that are expected to be well below occupational exposure limits; additional studies of this type are needed to improve exposure assessment. As occupation trends changed over time, it is also possible that RIDGE was unable to identify jobs with potential exposure because data about these jobs were not available in the training set. There are various occupations that were not represented in the past decade or two (e.g. firefighters), while some past jobs are not part of the current workforce. For other occupations, job tasks have expanded. Including an *a priori* decision rules approach can complement areas where the RIDGE might not perform at its best, although it depends on selecting the correct words, combination of words, and alternate words to describe a potentially exposed job. In reviewing job descriptions of jobs classified as false negatives, it is not always achievable to account for all possible keywords related to a specific job task or exposure. Additionally, if job description narratives have multiple misspellings, it can create a challenge for the RIDGE, since the RIDGE does not allow fuzzy matching and the misspelled word will not be eligible as a feature if it does not occur three or more times. Similarly, it can affect the *a priori* decision rules if the misspelled word is not included as a keyword. The negative effect of losing a word due to misspelling can often be compensated by other words in the narrative and if misspellings occur often enough, they remain as a feature with the appropriate context. Whenever feasible, the process of selecting keywords and occupational codes should involve more than one expert to increase the robustness of these decision rules. Since the machine-learning approach can be updated by adding new job description narratives in the training set to account for changes in the workforce over time, the prediction model can be updated at any time. We did not test the efficacy of these approaches in assigning additional exposure metrics, such as exposure intensity; it is possible that algorithms might perform more poorly for these more nuanced metrics.

Conclusions

This study adds to a body of literature that demonstrates the feasibility of developing a robust machine-learning prediction model for occupational exposure using a combination of known approaches for retrospective exposure assessment. Since the exposure scores—as classified by expert raters—were available, it also served as a learning tool to present the strengths and limitations of each individual approach, as well as the value of a combination of approaches.

Supplementary material

Supplementary data are available at *Annals of Work Exposures and Health* online.

Acknowledgments

The authors would like to thank all NBDPS collaborators and participants without whom this work would not be possible. This work was completed in partial fulfillment of the Doctor of Philosophy degree in Epidemiology in the Department of Environmental Health, Division of Epidemiology, University of Cincinnati College of Medicine through the support of Centers for Disease Control and Prevention (CDC) cooperative agreements under PA #96043, PA #02081, FOA #DD09-001, FOA #DD13-003, and NOFO #DD18-001 to the Centers for Birth Defects Research and Prevention participating in the National Birth Defects Prevention Study (NBDPS) and/or the Birth Defects Study To Evaluate Pregnancy exposureS (BD-STEPS), and by contract 200-2000-08018 from the CDC and the National Institute for Occupational Safety and Health; the Molecular Epidemiology in Children's Environmental Health pre-doctoral training grant (NIEHS T32-ES010957), and the Oak Ridge Institute for Science and Education (ORISE) fellowship. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Funding

This study was funded by the Centers for Disease Control and Prevention (CDC) cooperative agreements under PA #96043, PA #02081, FOA #DD09-001, FOA #DD13-003, and NOFO #DD18-001 to the Centers for Birth Defects Research and Prevention participating in the National Birth Defects Prevention Study (NBDPS) and/or the Birth Defects Study To Evaluate Pregnancy exposureS (BD-STEPS), and by contract 200-2000-08018 from the CDC and the National Institute for Occupational Safety and Health. This work was completed in partial fulfillment of the lead author's Doctor of Philosophy degree in Epidemiology in the Department of Environmental Health, Division of Epidemiology, University of Cincinnati College of Medicine through the support of the Molecular Epidemiology in Children's Environmental Health pre-doctoral training grant (NIEHS T32-ES010957), and the Oak Ridge Institute for Science and Education (ORISE) fellowship at CDC.

Conflict of interest statement

The authors do not have conflicts of interest to disclose.

References

Agency for Toxic Substances and Disease Registry. (1995) *Toxicological profile for polycyclic aromatic hydrocarbons (PAHs)*. Atlanta, GA: US Department of Health and Human Services, Public Health Service.

Alpaydin E. *Introduction to machine learning*: Cambridge, MA: MIT Press.

Bertke SJ, Meyers AR, Wurzelbacher SJ *et al*. (2012) Development and evaluation of a Naïve Bayesian model for coding causation of workers' compensation claims. *J Safety Res*; **43**: 327–32.

Bertke SJ, Meyers AR, Wurzelbacher SJ *et al*. (2016) Comparison of methods for auto-coding causation of injury narratives. *Accid Anal Prev*; **88**: 117–23.

Clavel J, Glass DC, Cordier S *et al*. (1993) Standardization in the retrospective evaluation by experts of occupational exposure to organic solvents in a population-based case-control study. *Int J Epidemiol*; **22**(Suppl 2): S121–6.

Fleming DA, Woskie SR, Jones JH *et al*. (2014) Retrospective assessment of exposure to chemicals for a microelectronics and business machine manufacturing facility. *J Occup Environ Hyg*; **11**: 292–305.

Florath I, Glass DC, Rhazi MS *et al*. (2019) Inter-rater agreement between exposure assessment using automatic algorithms and using experts. *Ann Work Expo Health*; **63**: 45–53.

Friesen MC, Locke SJ, Tornow C *et al*. (2014) Systematically extracting metal- and solvent-related occupational information from free-text responses to lifetime occupational history questionnaires. *Ann Occup Hyg*; **58**: 612–24.

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, Shortreed SM, Wheeler DC *et al*. (2015) 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. *Ann Occup Hyg*; **59**: 455–66.

Friesen MC, Wheeler DC, Vermeulen R *et al*. (2016) Combining decision rules from classification tree models and expert assessment to estimate occupational exposure to diesel exhaust for a case-control study. *Ann Occup Hyg*; **60**: 467–78.

Fritschi L. (2019) OccIDEAS – occupational exposure assessment in community-based studies. *Occup Med*; **69**: 156–57.

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.

Fritschi L, Siemiatycki J, Richardson L. (1996) Self-assessed versus expert-assessed occupational exposures. *Am J Epidemiol*; **144**: 521–7.

Lee DG, Lavoué J, Spinelli JJ *et al*. (2015) Statistical modeling of occupational exposure to polycyclic aromatic hydrocarbons using OSHA data. *J Occup Environ Hyg*; **12**: 729–42.

Lehto M, Marucci-Wellman H, Corns H. (2009) Bayesian methods: a useful tool for classifying injury narratives into cause groups. *Inj Prev*; **15**: 259–65.

Marucci-Wellman HR, Lehto MR, Corns HL. (2015) A practical tool for public health surveillance: Semi-automated coding of short injury narratives from large administrative databases using Naïve Bayes algorithms. *Accid Anal Prev*; **84**: 165–76.

Measure AC. (2014) *Automated coding of worker injury narratives*. Washington, DC: U.S. Bureau of Labor Statistics City.

Occupational Safety and Health Administration (OSHA). (2019) *OSHA annotated PELs Table Z-1*. Washington, DC: OSHA.

Peters S, Glass DC, 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.

Reefhuis J, Gilboa SM, Anderka M *et al*; National Birth Defects Prevention Study. (2015) The national birth defects prevention study: a review of the methods. *Birth Defects Res A Clin Mol Teratol*; **103**: 656–69.

Rezagholi M, Mathiassen SE. (2010) Cost-efficient design of occupational exposure assessment strategies–a review. *Ann Occup Hyg*; **54**: 858–68.

Rocheleau CM, Lawson CC, Waters MA *et al*. (2011) Inter-rater reliability of assessed prenatal maternal occupational exposures to solvents, polycyclic aromatic hydrocarbons, and heavy metals. *J Occup Environ Hyg*; **8**: 718–28.

Sauvé JF, Lavoué J, Nadon L *et al*. (2019) A hybrid expert approach for retrospective assessment of occupational exposures in a population-based case-control study of cancer. *Environ Health*; **18**: 14.

Teschke K, Olshan AF, Daniels JL *et al*. (2002) Occupational exposure assessment in case-control studies: opportunities for improvement. *Occup Environ Med*; **59**: 575–93; discussion 594.

Tinker SC, Carmichael SL, Anderka M *et al*; Birth Defects Study To Evaluate Pregnancy exposureS. (2015) Next steps for birth defects research and prevention: The birth defects study to evaluate pregnancy exposures (BD-STEPS). *Birth Defects Res A Clin Mol Teratol*; **103**: 733–40.

U.S. Environmental Protection Agency. (2008) *Polycyclic aromatic hydrocarbons (PAHs)*.

Vallmuur K. (2015) Machine learning approaches to analysing textual injury surveillance data: a systematic review. *Accid Anal Prev*; **79**: 41–9.

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.

White E, Armstrong B, Saracci R. (2008) *Principles of exposure measurement in epidemiology*. Oxford University Press: New York, NY.

Williamson A, Feyer AM, Stout N *et al*. (2001) Use of narrative analysis for comparisons of the causes of fatal accidents in three countries: New Zealand, Australia, and the United States. *Inj Prev*; **7**(Suppl 1): i15–20.

Yoon PW, Rasmussen SA, Lynberg MC *et al*. (2001) The National Birth Defects Prevention Study. *Public Health Rep*; **116**(Suppl 1): 32–40.