Annals of Work Exposures and Health, 2020, Vol. 64, No. 1, 13–24
doi: 10.1093/annweh/wxz077
Advance Access publication 31 October 2019
Original Article

Original Article

Bayesian Hierarchical Modelling of Individual Expert Assessments in the Development of a General-Population Job-Exposure Matrix

Jean-François Sauvé[1,2], Marie-Pierre Sylvestre[2,3], Marie-Élise Parent[2,3,4] and Jérôme Lavoué[1,2*]

[1]Department of Environmental and Occupational Health, School of Public Health, Université de Montréal, 2375 chemin de la Côte Ste-Catherine, Montréal, Québec H3T 1A8, Canada; [2]Centre de recherche du CHUM, 850 rue St-Denis, Montréal, Québec H2X 0A9, Canada; [3]Department of Social and Preventive Medicine, School of Public Health, Université de Montréal, 7101 Avenue du Parc, Montréal, Québec H3N 1X9, Canada; [4]INRS-Institut Armand-Frappier, Université du Québec, 531 Boul. des Prairies, Laval, Québec, H7V 1B7, Canada

*Author to whom correspondence should be addressed. Tel: +1-514 343-6111 #3108; fax: +1-514-343-2200; e-mail: jerome.lavoue@umontreal.ca

Submitted 28 November 2018; revised 26 June 2019; editorial decision 5 September 2019; revised version accepted 14 September 2019.

Abstract

The CANJEM job-exposure matrix compiles expert evaluations of 31 673 jobs from four population-based case–control studies conducted in Montreal. For each job, experts had derived indices of intensity, frequency, and probability of exposure to 258 agents. CANJEM summarizes the exposures assigned to jobs into cells defined by occupation/industry, agent, and period. Some cells may, however, be less populated than others, resulting in uncertain estimates. We developed a modelling framework to refine the estimates of sparse cells by drawing on information available in adjacent cells. Bayesian hierarchical logistic and linear models were used to estimate the probability of exposure and the geometric mean (GM) of frequency-weighted intensity (FWI) of cells, respectively. The hierarchy followed the Canadian Classification and Dictionary of Occupations (CCDO) classification structure, allowing for exposure estimates to be provided across occupations (seven-digit code), unit groups (four-digit code), and minor groups (three-digit code). The models were applied to metallic dust, formaldehyde, wood dust, silica, and benzene, and four periods, adjusting for the study from which jobs were evaluated. The models provided estimates of probability and FWI for all cells that pulled the sparsely populated cells towards the average of the higher-level group. In comparisons stratified by cell sample size, shrinkage of the estimates towards the group mean was marked below 5 jobs/cell, moderate from 5 to 9 jobs/cell, and negligible at ≥10 jobs/cell. The modelled probability of three-digit cells were slightly smaller than their descriptive estimates. No systematic trend in between-study differences in exposure emerged. Overall, the modelling framework for FWI appears to be a suitable approach to refine CANJEM estimates. For probability, the models could be improved by methods better adapted to the large number of cells with no exposure.

Keywords: retrospective occupational exposure assessment; job-exposure matrix; expert assessment; hierarchical model

Introduction

The CANJEM project aimed at developing a general-population job-exposure matrix (JEM) from the information of >30 000 jobs evaluated by experts during four population-based case–control studies of cancer conducted in Montreal, Canada since the 1980s (Sauvé *et al.*, 2018; Siemiatycki and Lavoué, 2018). In this process, the exposures assigned to individual jobs (where each job represents an occupation held by a subject for at least 6 months) for 258 chemical and physical agents were summarized in CANJEM into strata of occupation or industry (available in several standardized classifications) and employment period. Both the occupation/industry and the period dimensions are defined over several levels of resolution, from specific occupations/industries to broader categories, or from a single global period to a stratification into four shorter periods. Each cell provides a descriptive summary of the exposures assigned to jobs according to their probability, confidence, intensity, frequency, and frequency-weighted intensity (FWI). CANJEM has been recently used to estimate the risk of thyroid cancer associated with exposure to pesticides and biocides (Zeng *et al.*, 2017) and to estimate the prevalence of occupational exposure to 21 chemicals in the Northwestern United States (Doubleday *et al.*, 2019).

CANJEM is akin to a JEM developed from a database of workplace measurements as it is based on a finite sample of jobs held by subjects in the four studies conducted over a 25-year period. The information for each job represents expert judgment on the presence and ordinal level of exposure to a series of agents over the time period covered by the job, as opposed to a single objective measurement of concentration in air. The number of jobs available to develop exposure estimates thus varies across cells, being lower for less prevalent occupations or industries in the population. Increasing the resolution of the occupation/industry groups and periods also implies that the finite set of jobs is distributed across a larger number of categories, thereby further decreasing sample size per cell. One way to obtain a more precise estimate of exposure for a cell based on few jobs (e.g. snowmobile repairers) is to use the estimate of a broader occupation group (e.g. motor vehicle mechanics), pooling jobs across the nested occupations. This may represent a useful approach when the exposure profile in one occupation (or industry) is comparable to the other occupations within the same group. On the other hand, this could introduce bias if the exposure profile of the broader group is not indicative of the exposure in a specific occupation. Hierarchical models represent an alternative approach that could provide a compromise between the unbiased, but less precise information of cells at finer resolution, and the more precise, but also potentially biased information of coarser resolutions. The use of these models structured by the occupation/industry systems allows for cells based on a few data points to draw information from other, more populated cells associated with similar occupations within a broader group.

Hierarchical models have been applied to the interpretation of workplace measurement data for purposes of comparisons with exposure limits (Banerjee *et al.*, 2014) and in occupational epidemiology to account for similarities in exposure profiles among workers, job titles, or facilities (Friesen *et al.*, 2006; Portengen *et al.*, 2016; Toti *et al.*, 2006). Other examples include combining a generic JEM with measurement data to estimate quantitative exposure levels by occupation, where the occupations were grouped by their categorical JEM rating (Peters *et al.*, 2011, 2016; Friesen *et al.*, 2012). Hierarchical models have also been used in the evaluation of lung cancer risk for 184 agents by pooling information across agents sharing similar chemical characteristics and/or prior evidence of carcinogenicity (Momoli *et al.*, 2010). More recently, Roberts *et al.* (2018) used Bayesian hierarchical models based on the structure an occupational classification to impute exposure estimates in the development of a general-population JEM from a large database of noise measurements. Compared to similar frequentist approaches, Bayesian inference is more easily amenable to modelling complex multilevel structures (Gelman and Hill, 2007) and allows incorporating prior knowledge on the distribution of the parameter(s) of interest.

In this article, we developed a Bayesian modelling framework applied to the expert ratings to refine the estimates of sparse CANJEM cells by building on the information contained in cells of similar occupations. The application of models also provided an opportunity to explore trends in exposure differentiated by study and to compare the predicted exposure estimates of cells to those obtained from descriptive summaries.

Methods

The Montreal case–control studies

Study populations

CANJEM is based on data from four population-based case–control studies in Montréal, Canada. Study 1 (conducted 1979–1986) investigated 19 different sites of cancer among men aged 35–90 years (3726 cancer patients and 533 population controls) (Siemiatycki *et al.*, 1987). Study 2 (1996–2001) was a study of lung cancer and included males and females aged 35–75 years (1203 cases and 1513 population controls) (Pintos *et al.*, 2012). Study 3 (1996–1997) was a study of breast cancer and included women aged 50–75 years (608 cases and 667 cancer controls) (Labrèche *et al.*, 2010). Study 4 (2000–2004) was a study of glioma and meningioma tumours and represented the Quebec and Ontario portions of the multi-centric INTEROCC study (Lacourt *et al.*, 2013), and included men and women aged between 30 and 59 years of age (218 cases and 414 population controls).

Occupational exposure assessment

The expert approach to exposure assessment described in Gérin *et al.* (1985) was developed during Study 1 and applied in subsequent studies. Briefly, complete occupational histories including job titles, employment duration, tasks performed, work environment and conditions, and product and equipment use were collected from questionnaires and extensive face-to-face interviews with subjects, or proxy respondents. A team of trained chemists and industrial hygienists reviewed each job description, blind to the subject's case/control status, to assign standardized job and industry titles and to assess exposures to a predefined list containing approximately 300 chemical physical and biological agents. Exposure was rated by its intensity (low, medium, high), its frequency (hours per week), and the experts' level of confidence in the assessment (possible, probable, definite). Jobs judged exposed to an agent at a concentration equivalent to or less than a background non-occupational level were considered unexposed.

Exposure information in CANJEM

The occupational histories and exposure data associated with 31 673 jobs served as the foundation of CANJEM, with 15 067 jobs from Study 1 (47.6%), 10 371 from Study 2 (32.7%), 3510 from Study 3 (11.1%), and 2725 from Study 4 (8.6%). CANJEM summarizes the exposure information of these jobs into three dimensions: agents, occupations/industries, and periods. The agent axis includes 258 agents. The occupation/industry dimension is available in seven standard classification schemes. For each agent, exposure estimates can be obtained at several resolution levels of the selected classification, from broader groupings (e.g. service occupations) to the most detailed categories (e.g. waiters). The period dimension is available in three levels of resolution: a single global period (1930–2005), two periods (1930–1969 and 1970–2005), and four periods (1930–1949, 1950–1969, 1970–1984, and 1985–2005). These periods were defined *a priori* based on broad population-level changes potentially affecting exposure (e.g. changes in regulatory environment in the 1970s). CANJEM can be consulted freely at www.canjem.ca.

The exposure profile of jobs in each cell is represented by five indices: probability, confidence, intensity, frequency, and FWI of exposure. Probability represents the proportion of jobs exposed among all jobs in the cell. Confidence is the relative proportion of jobs with possible, probable, and definite ratings. Similarly, intensity of exposure presents the relative proportion of exposed jobs across the low, medium and high ratings, and frequency the relative proportion of jobs exposed <2 h, 2 to <12 h, 12 to <40 h, and 40 h per week or more. Last, the continuous FWI index represents the intensity of exposure averaged over a 40-h workweek, computed by multiplying the intensity ratings with the frequency of exposure relative to a baseline of 40 h. In computing FWI, weights of 1, 5, and 25 were assigned to the low-, medium-, and high-intensity categories, respectively (Lavoué *et al.*, 2012; Sauvé *et al.*, 2018). FWI in CANJEM cells is represented as the median value across exposed jobs.

Model development

General framework

We used hierarchical models based on the structure of the occupational/industrial classification, in which one model provided exposure estimates for all cells across all levels of a classification system. Cells from each period were modelled separately for two reasons. First, we wanted to allow for the probability or FWI of cells to vary between periods for the same occupation. Second, jobs could span several years and belong to several contiguous periods, which represented a challenge to the use of a single model applied to several periods at once. Therefore, for one combination of agent, period, and exposure index, a single model provides estimates for all categories of the selected occupational or industrial classification across all resolutions.

Furthermore, although all studies relied on the same general data collection and exposure assessment framework, shifts in the definition of exposure indices, refinement of questionnaires, and accrual of experience may have caused differences in exposure estimates for a comparable situation. This would have been the most important for the time gap between the first study conducted by the group in the early 1980s, and the other studies that were conducted some 15 years later. To account for potential shifts in exposure coding, the models included a binary indicator separating the older Multisite study from the others.

Development and structure of models

The models were developed for two exposure indices, probability and FWI, and applied to CANJEM defined by four periods and the 1971 Canadian Classification and Dictionary of Occupations (CCDO) (Department of Employment and Immigration, 1971). This classification is structured with four hierarchical levels: two-digit major groups, three-digit minor groups, four-digit unit groups, and seven-digit occupations, the latter featuring 7907 unique codes. Only the three-, four-, and seven-digit levels were included in the models since the two-digit major group strata was considered too broad. The models were applied to five agents (metallic dust, formaldehyde, wood dust, crystalline silica, and benzene) to encompass a diversity of physical forms. All available cells in CANJEM were used in the modelling, without restriction on sample size, because cells based on a single job could still provide exposure information for higher-level groups.

Because of the expectations of small sample size, empty cells, and uneven distribution of data, we chose to develop the models using Bayesian inference, which is also naturally suited to the application of hierarchical models. General introductions to Bayesian inference can be found in Gelman *et al.* (2014) and Kruschke (2015), with more detailed information on computational methods provided in Carlin and Louis (2009). Recent applications of Bayesian methods in assessing occupational exposures have been reviewed by Ramachandran (2019).

The model for FWI was applied to the individual exposed jobs separately for each agent and period. Linear models were applied to the log-transformed FWI values to estimate the geometric mean (GM) FWI by occupational group based on the structure shown below.

$$\ln(\mathrm{FWI}_{hijk}) = \beta^{(\mathrm{FWI})}_{\mathrm{Study}} + \beta^{(\mathrm{FWI})}_{3d_h} + b^{(\mathrm{FWI})}_{4d_{hi}} + b^{(\mathrm{FWI})}_{7d_{hij}} \qquad (1)$$

where $\ln(\mathrm{FWI}_{hijk})$ is the log-transformed FWI value of the *k*th job in the *j*th seven-digit group in the *i*th four-digit group in the *h*th three-digit group.

The three-digit groups were entered as fixed effects in the model. The four-digit groups ($b_{4d_{hi}}$) in equation (1) were entered as a random-effects nested within three-digit groups (β_{3d_h}) and the seven-digit groups ($b_{7d_{hij}}$) were nested within the four-digit groups. Last, the term for study was entered as a binary indicator in the models for the two middle periods only (1950–1969 and 1970–1984) where the job histories overlapped the most. This term was excluded for the other periods since 73% of jobs in the period 1930–1949 came from Multisite and 95% of jobs for 1985–2005 came from the more recent studies.

For probability of exposure, which is expressed as a proportion, a logistic model was used along with the same core structure used to model FWI. The number of exposed jobs in a cell (*Nexp*) was modelled as a binomial distribution defined by the proportion of jobs exposed (π) and the total number of jobs evaluated (*Ntot*) [equation (2)]. The logit of π was assumed to follow a normal distribution based on the mean of the seven-digit group and the level of the study variable (when applicable).

$$Nexp_{hijk} \sim \mathrm{Binomial}(\pi_{hijk},\ Ntot_{hijk}) \qquad (2)$$

where $Nexp_{hijk}$ represents the number of exposed jobs in the cell for the *j*th seven-digit occupation and the *k*th study. The logit of π was then modelled according to equation (3), allowing for the concurrent estimation of probability across the three-, four-, and seven-digit groups.

$$\mathrm{logit}(\pi_{hijk}) = \beta^{(\pi)}_{Study} + \beta^{(\pi)}_{3d_h} + b^{(\pi)}_{4d_{hi}} + b^{(\pi)}_{7d_{hij}} \qquad (3)$$

where π_{hijk} represents the estimated proportion of exposed job of the *k*th study in the *j*th seven-digit group in the *i*th four-digit group in the *h*th three-digit group.

Implementation of the Bayesian models

We fit the Bayesian models using the JAGS 3.4.0 software (Plummer, 2003). The JAGS code is provided in Supplementary Material (available at *Annals of Work Exposures and Health* online).

Priors for the coefficients of three-digit groups (β_{3d}) and study were normal distributions with mean 0 and variance 1000. Priors for the between-occupation and between-unit group variances, and for the within-occupation variance (FWI only), were uniform distributions on the scale of the standard deviation bounded between 0.001 and 100. Each model for FWI was fitted using 12 Markov chain Monte Carlo (MCMC) chains with 75 000 iterations each, discarding the first 25 000 iterations for burn-in, for a total of 600 000 iterations used for inference. Since the models for probability applied to all CANJEM cells (not only those with exposed jobs), we used a larger number of iterations (275 000 per chain), discarding the first 25 000 iterations and keeping the results of one out of each 10 iteration (total of 300 000 iterations kept). Convergence was assessed using the Brooks–Gelman–Rubin statistic, or $\hat{R}$ (Brooks and Gelman, 1998), where a value close to 1 indicates convergence for a given parameter. For FWI, convergence was reached for all parameters with $\hat{R}$ lower than 1.1 (Gelman and Shirley, 2011) for all combinations of agents and periods. For probability, there remained on average 1% of the model parameters with Rhat values above 1.1 (range across combinations of period and agent = 0–6%).

Development of predictions for CANJEM cells

The hierarchical model structure allowed for predictions to be made for the probability of exposure or the GM of FWI for all cells across the three levels of the CCDO classification for one combination of agent and period. One important feature of hierarchical models is the borrowing of information across the data by shrinking the more imprecise estimates in the direction of the broader group. Exposure estimates for cells with few observations will tend to be pulled more heavily towards the mean of the higher-level group, more so when their (unshrunk) estimates differ markedly from the group mean (Gelman and Hill, 2007). On the other hand, cells with more observations would be less affected. This shrinkage allows an increase in precision in estimates while applying some level of bias towards the estimate of the higher-level group (Greenland, 2000). The relative influence of the different occupations in the models was mostly driven by the distribution of the data. Recently, Quick *et al.* (2017) proposed a method to modulate the relative influence of the exposure groups based on sample sizes in constructing informative prior distributions. However, we did not use this approach because of the large number of occupations distributed across three hierarchical levels in our models.

As an illustration consider a group of motor vehicle mechanics as one level, with two subgroups forming the lower level: automobile mechanics, with a FWI value for diesel exhaust of 0.5 based on 50 jobs (equivalent to 20 h exposed at low intensity per week), and heavy truck mechanics, with a single job with a FWI of 25 (i.e. 40 h at high intensity). The resulting overall estimate for mechanics would therefore be mainly based on jobs from automobile mechanics. Provided this distribution of mechanics jobs reflects the distribution in the base population (i.e. higher proportion of automobile mechanics), the estimate would accurately represent the overall exposure profile for 'mechanics'. However, the estimate for truck mechanics would be pulled towards the overall estimate for mechanics, i.e., closer to 0.5 than 25, itself driven by automobile mechanics. If an FWI of 25 accurately reflects the exposure of truck drivers in the population, this pulling effect is undesirable. On the other hand, if the exposure of truck drivers in the population is actually lower and the very high exposure for the single observation available in our database is due to random sampling, the pooling of the data would provide a better estimate.

Because it is not possible to discriminate between these two scenarios for every situation that may arise in CANJEM, we conducted an evaluation of the impact of sample size on the robustness of the estimates to shrinkage. This was done with the aim of finding a compromise value allowing for some, but not extreme, shrinkage for using the results of a cell.

The inclusion of a variable for the source study of jobs in the models allowed for predictions to be made for a cell corresponding to a case reflecting only the Multisite study, only the site-specific studies, or a combination of both. In consultation with the experts, we made the predictions for CANJEM cells to reflect a situation where 75% of the information came from the site-specific studies, and the remaining 25% from the earlier Multisite study. We gave a higher weight to the more recent site-specific studies because experts had more experience with the coding approach and had access to a larger pool of information to reconstruct past exposures. We adjusted the predictions for the periods where the study term was omitted from the models. In those cases, we used the median of the posterior distribution for the study parameter from the model of the nearest period. Details for the adjustment for study on the predictions are presented in Supplementary Appendix (are available at *Annals of Work Exposures and Health* online).

Results

Descriptive statistics of the exposure data

The total number of jobs available, the number of cells by CCDO level, and the number of exposed jobs by agent for each of the four periods are presented in Table 1. The total is greater than the total number of jobs (31 673) because jobs could be included in more than one period. All of the 81 three-digit minor groups had at least one job in the two middle periods. However, no jobs were available in period 1930–1949 for minor

Table 1. Total number of jobs per time period and corresponding number of exposed jobs by agent

	Time period			
	1930–1949	1950–1969	1970–1984	1985–2005
Overall				
Number of jobs available[a]	9444	17 147	13 450	6405
Number of seven-digit occupations[b]	1743	2408	2082	1289
Number of four-digit unit groups[b]	436	469	461	392
Number of three-digit minor groups[b]	79	81	81	80
Number of exposed jobs by agent (%)				
Metallic dust	1258 (13.3%)[c]	2065 (12.0%)	1344 (10.0%)	402 (6.3%)
Formaldehyde	965 (10.2%)	1960 (11.4%)	1522 (11.3%)	663 (10.4%)
Wood dust	1007 (10.7%)	1525 (8.9%)	916 (6.8%)	341 (5.3%)
Silica	807 (8.5%)	1480 (8.6%)	907 (6.7%)	277 (4.3%)
Benzene	571 (6.0%)	1055 (6.2%)	541 (4.0%)	145 (2.3%)

[a] Because jobs could be present in two or more adjacent periods, the total is greater than the number of jobs used in the construction of CANJEM (n = 31 673).

[b] Number of groups with at least one job available.

[c] Percentage of exposed jobs relative to the total number of jobs assessed within the time period.

groups 235 (occupations in library, museum, and archival sciences) and 239 (other occupations in social sciences), and for minor group 731 (Fishing, trapping and related occupations) in period 1985–2005.

Metallic dust had the highest proportion of exposed jobs for the first two periods, while formaldehyde ranked highest for the two most recent periods. The number of seven-digit cells with at least one exposed job varied from 94 (benzene, 1985–2005) to 716 (metallic dust, 1950–1969).

Modelling

Between-study differences in exposure

The associations between site-specific studies and the probability and GM of FWI in cells relative to those of the Multisite study are presented in Table 2. The between-study differences in probability of exposure were generally small among the combinations of agents and periods, with odds ratios (ORs) close to 1. The largest difference was observed for benzene in the period 1950–1969 where jobs from site-specific studies were twice likely to be exposed compared to those from the Multisite study (OR = 1.93, 90% confidence interval = 1.66–2.24). The influence of site-specific studies on the GM of FWI of cells (Table 2) was expressed as a ratio relative to a reference of 1 for Multisite. Site-specific studies were associated with FWI levels on average 0.75–0.80 of those in Multisite for silica and 0.50 for wood dust and metallic dust. In the case of formaldehyde and benzene, the FWI levels of jobs were comparable between the two study groups.

Predicted probability and GM of FWI of cells

To illustrate the distribution of the information on exposure in cells across the levels of the classification, Fig. 1 presents the predicted probability and GMs of FWI for exposure to formaldehyde (where the between-study differences were negligible) among cells nested in the minor group of fabricating, assembling and repairing occupations, wood products (CCDO 854) for the period 1970–1984. Approximately half of all jobs were associated with the occupation of cabinetmakers (CCDO 8541-110), while most of the other occupations were based on one or two jobs.

Effect of sample size on the sensitivity to shrinkage

Among the occupations listed in Fig. 1, the predicted FWI of Laminating-press tenders was heavily pulled towards the overall mean due to its small sample size and large value relative to the other occupations. Another illustration of the influence of cell sample size on shrinkage of the estimates is presented in Fig. 2, which compares the observed to the predicted estimates of exposure to formaldehyde among all seven-digit cells within the minor group 855/856 (Fabricating, assembling, and repairing occupations: Textile, fur, and leather products) for the period 1950–1969. The 62 seven-digit cells were categorized in three groups by cell sample size (exposed jobs for FWI): fewer than 5 jobs, 5–9 jobs, and 10 jobs or more. For both probability and FWI, the difference between the predicted and the observed estimates decreases from the leftmost panel (<5 jobs) to the rightmost one (≥10 jobs), with a marked reduction in sensitivity with a sample size of at least 5

Table 2. Relative influence of site-specific studies on the probability and FWI of exposure of cells relative to Multisite, stratified by time period and agent

	Probability		Frequency-weighted intensity	
	Odds ratio (90% CI) (reference: multisite = 1)		Geometric mean ratio (90% CI) (reference: multisite = 1)	
Period[a]	1950–1969	1970–1984	1950–1969	1970–1984
Metallic dust	1.09 (0.94–1.26)[b]	0.95 (0.80–1.13)	0.61 (0.56–0.68)[c]	0.56 (0.50–0.64)
Formaldehyde	0.93 (0.83–1.05)	1.05 (0.92–1.21)	1.18 (1.09–1.28)	1.08 (0.98–1.18)
Wood dust	1.27 (1.10–1.47)	1.19 (1.00–1.42)	0.51 (0.46–0.58)	0.45 (0.38–0.52)
Silica	0.72 (0.62–0.84)	0.93 (0.78–1.11)	0.72 (0.64–0.82)	0.80 (0.68–0.94)
Benzene	1.93 (1.66–2.24)	1.17 (0.96–1.43)	1.02 (0.88–1.17)	1.05 (0.87–1.26)

[a] The inclusion of study in the models was limited to periods 1950–1969 and 1970–1984.

[b] Odds ratio and 90% credible interval for site-specific studies, relative to a reference of 1 for the Multisite study.

[c] Ratio between the GM of FWI of cells from site-specific studies relative to a reference of 1 for the GM of cells from Multisite study, computed as $\exp(\beta_{study})$. The 90% credible interval around the ratio of the GMs is in parentheses.

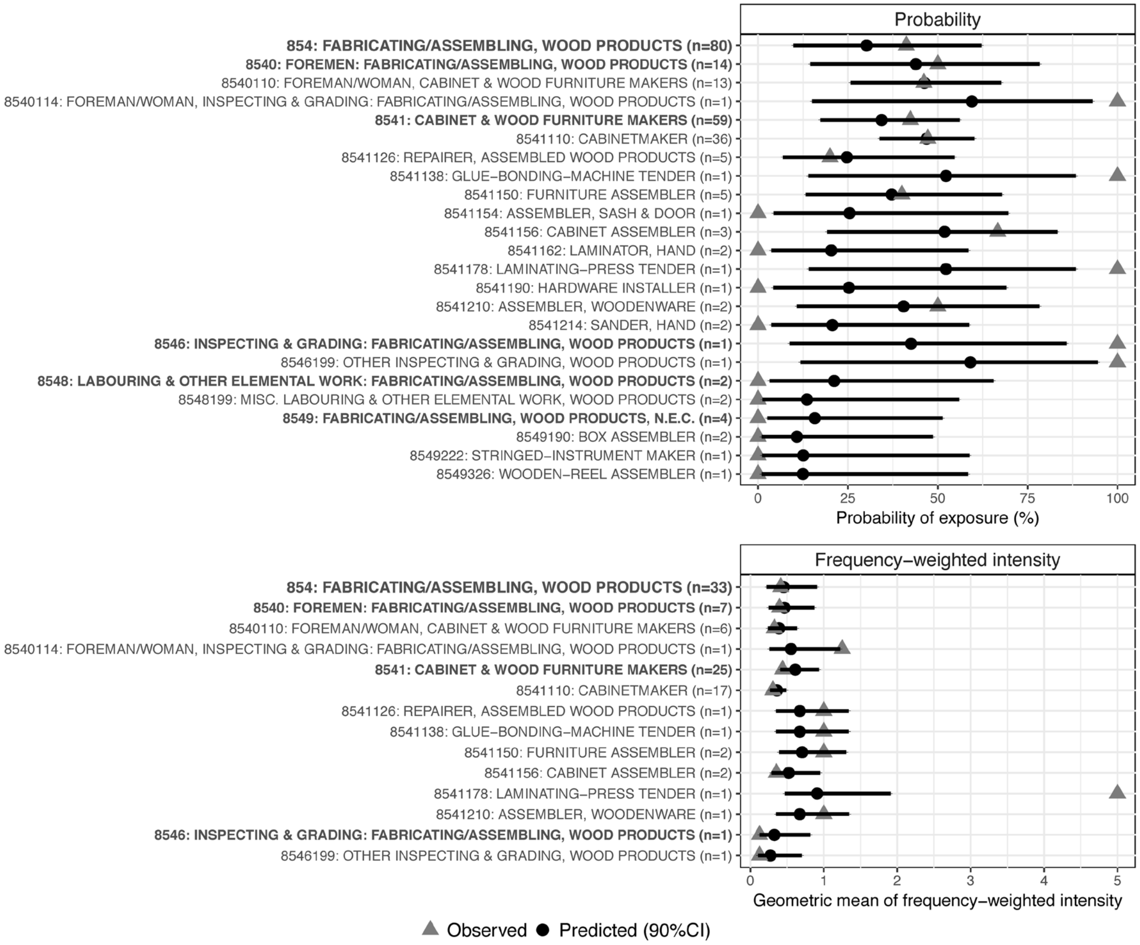

Figure 1. Comparison of the observed and predicted probability and GM for FWI for exposure to formaldehyde among cells nested in CCDO minor group 854 (Fabricating, assembling and repairing occupations, wood products), period 1970–1984.

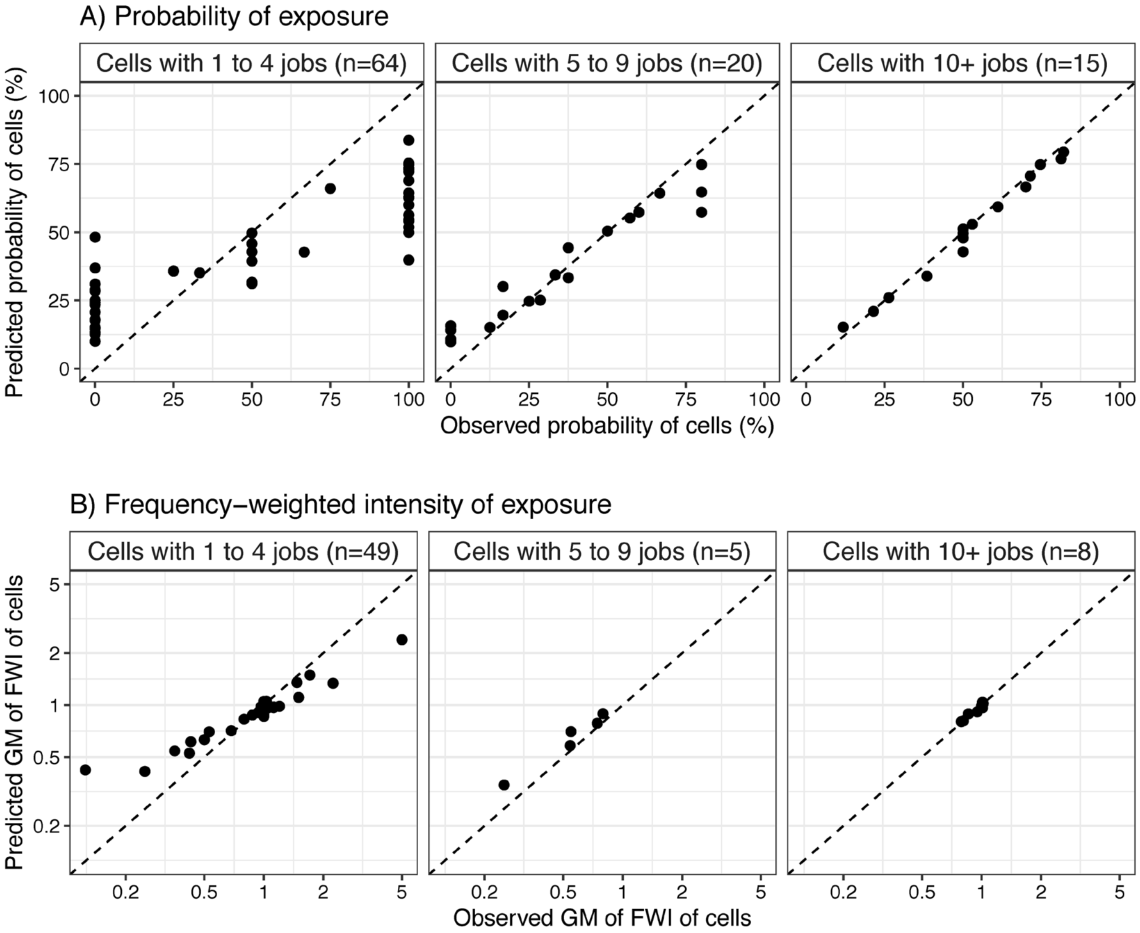

Figure 2. Comparison of the observed and predicted probability and FWI of exposure to formaldehyde, stratified by cell sample size. Cells shown are all seven-digit cells (*n* = 99 for probability, *n* = 62 for FWI) within major group 855/856 (Fabricating, assembling and repairing occupations: Textile, fur and leather products) in period 1950–1969.

jobs. The pattern observed in Fig. 2 was generally consistent across other combinations of three-digit groups, agents, and time periods (results not shown).

Predicted versus observed probability and FWI

Figure 3 presents a comparison of the distribution of the observed and predicted probability (Fig. 3a) and FWI (Fig. 3b) of all cells stratified by CCDO level, using exposure to formaldehyde in period 1970–1984 as an illustration. For FWI, the models pulled the very low or very high estimates towards the overall average for cells at the seven-digit occupation and four-digit unit group levels, where no systematic differences were observed in one direction or another. A different pattern emerged for probability where shrinkage for the three-digit cells went systematically in the direction of lower probability of exposure, with a median decrease of 1.7% in the predicted probability of cells relative to their observed values (interquartile interval 0–6.4%). This pattern was consistent throughout the analyses, where the median decrease in the predicted probability of three-digit cells ranged 0–4.3% (median 0.8%) across the agents and periods.

Discussion

We developed Bayesian models to refine the probability and FWI estimates of CANJEM cells that were based on a small number of jobs by borrowing information on exposure available in other related occupations using the structure of a standardized classification. The resulting estimates are a compromise between the level of information on exposure specific to jobs evaluated in a cell, and the information available in other cells within the same broader occupational group.

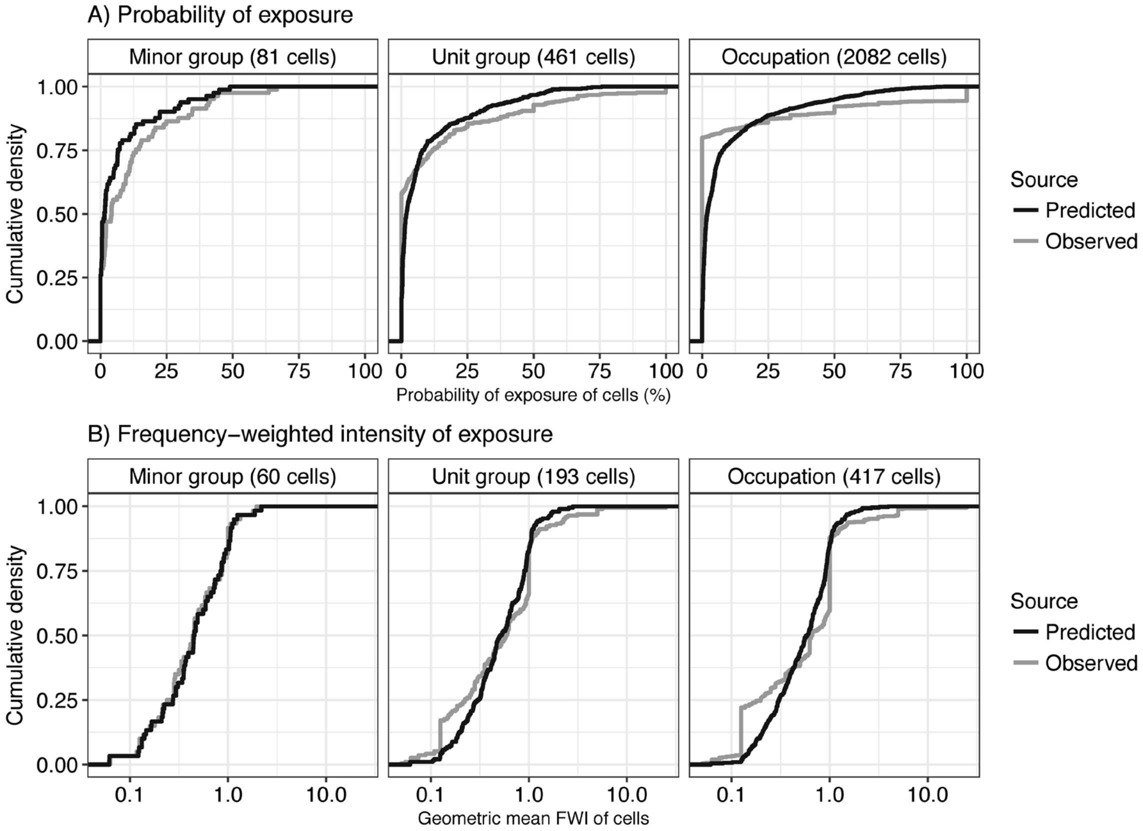

Figure 3. Comparison of empirical cumulative distribution functions of the observed and predicted estimates of cells across all three levels of the CCDO classification, for the probability and FWI of exposure to formaldehyde, period 1970–1984.

Predicted probability and FWI

Effect of sample size and classification structure on shrinkage

The borrowing of information on exposure between cells in the models was organized by the hierarchical structure of the occupational classification, which implies some level of exchangeability in exposure across occupations. For example, this would suppose that the exposure to formaldehyde is *a priori* comparable between the various occupations of cabinet and wood furniture makers in the absence of exposure data specific to an occupation. The estimates of individual cells would then draw on the information available within the larger pool of cabinet and wood furniture makers in the models, which would increase the precision by adding some amount of bias compared with a purely descriptive estimate. The trade-off is greater for cells with an outlying estimate relative to the other cells and with a smaller sample size. This shrinkage property can be useful in facilitating the inclusion of groups with a small sample size in an analysis that could otherwise result in unstable estimate. However, the possibility a large bias outweighing the increased precision of an estimate was a concern in the context of CANJEM due to the challenge of differentiating the lack of compatibility in the exposures between occupations from random variation for a large number of agents and occupational and industrial classifications. This challenge was also found with a quantitative JEM for noise (Roberts *et al.*, 2018) where some managerial occupations with *a priori* low exposure had high noise exposure predicted by the model due to borrowing information from industrial production and agricultural managers.

The evaluation of shrinkage showed that overall, cells with fewer than five jobs for probability, or exposed jobs for FWI, tended to be quite sensitive to the shrinkage effect (Fig. 2), while cells with five to nine jobs were more robust, and shrinkage was negligible for those with at least 10 data points. A sample size of five jobs, while an arbitrary threshold, may represent a reasonable starting point in using the estimate of a cell that accounts to some extent for the information available in similar occupations, without being overly sensitive to shrinkage towards the group mean. While the structure of the classification may not always be representative of the distribution in exposures, the impact of potential misspecification is therefore limited when using a sample size of at least five jobs per cell and avoids defining alternative schemes to group occupations based on exposures.

Overall trends in the predicted probability and FWI of cells

The distribution of the predicted GMs of FWI of cells showed that the more outlying estimates (high or low) were pulled towards the mean, suggesting the suitability of the models in pooling the exposure information across CANJEM cells. However, the model for probability resulted in a tendency to predict lower values at the highest hierarchical level (three-digit groups). This trend may be due to the application of logistic models to a distribution of (seven-digit) cells for which between 74 and 93% had no exposed job, depending on the agent and period. The evaluation of MCMC history plots also showed issues of convergence in cells with no exposed jobs compared with those with exposed jobs (results not shown). We also explored alternative models (see Supplementary Material, available at *Annals of Work Exposures and Health* online) such as linear models which, while allowing predictions outside of the 0–100% range, resulted in a distribution of predicted probability closer to FWI in Fig. 3. While the amplitude of the systematic shift towards lower probabilities was limited, the use of models adapted for zero counts, such as zero-inflated binomial regression in a hierarchical framework (Hall, 2000), might constitute an improved strategy to model the probability of exposure. The adaptation of these models to the multiple hierarchical levels of the classifications and the unbalanced structure of the data would however require further methodological development. Further developments could also include models adapted to the ordinal indices of intensity and frequency of exposure. However, the large proportion of empty cell counts also represents a challenge to the application of traditional modelling approaches.

Between-study differences in exposure

Overall, the between-study differences were generally small. The only exceptions were the lower FWI for wood dust and metallic dust, and higher probability of exposure for benzene (1950–1969 only) in site-specific studies. The differences between studies could be due to the increased experience and familiarity of the team with the exposure assessment method over time, and changes in the meaning benchmarks for the intensity categories and for the background environmental exposure level (Pintos *et al.*, 2012).

The inclusion of a variable for the source studies in the model also allowed us to weigh the relative influence of each study across all cells in the predictions. In contrast, the influence of the study on exposure estimates obtained using a descriptive approach might vary from cell to cell depending on the distribution of the jobs between the two study levels within each cell.

Extension of the models

In addition to extending the models to the ordinal indices of intensity and frequency, potential developments could also made to modify the hierarchical structure of the models to allow for borrowing information across periods. The addition of period in a cross-classified hierarchical design would permit information for a cell to be drawn from nearby occupation as well as from nearby periods (Browne *et al.*, 2001; Jones and Burstyn, 2016). The issue of jobs belonging to more than one period would however remain a challenge for this development.

Conclusion

The models applied to the index of FWI appeared to have adequately weighted the influence of cells between the hierarchical levels on the final estimates. Their application to probability was however suboptimal, likely due to the considerable number of cells with no exposure. The framework presented here can be useful in developing sources of exposure information from existing data sets of measurements or expert evaluations for which job or industry titles are available and where the issue of spare data may arise.

Supplementary Data

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

Funding

The development of CANJEM was funded by the Société de recherche sur le cancer (SRC) through the GrePEC initiative (grant no. 16264). The case-control studies were funded by several agencies, including the National Health Research and Development Programme, Health Canada, the National Cancer Institute of Canada, the Medical Research Council of Canada, the Canadian Institutes for Health Research, and the National Institutes of Health (NIH). Some of the computations were conducted on the supercomputer Briarée at the Université de Montréal, managed by Calcul Québec and Compute Canada. J.F.S. was supported by doctoral training scholarships from the Fonds de recherche du Québec—Santé (FRQS) and the Institut de Recherche Robert-Sauvé en Santé et en Sécurité du Travail. M.E.P. and J.L. had a salary award from the FRQS. M.P.S. holds a chercheur-boursier Junior 1 award from the FRQS. The authors declare no other conflict of interest relating to the material presented in this article.

Disclaimer

The authors have no conflict of interest to declare.

Acknowledgements

The original approach to expert assessment of exposure from detailed job histories was developed primarily by Michel Gérin and Jack Siemiatycki, with input throughout from the various teams responsible for exposure assessment for the original studies. Jack Siemiatycki was the principal investigator of the multi-site and lung cancer studies, and is the co-leader of CANJEM. Many chemists and hygienists have contributed their expertise to the coding of the previous studies on which this method was based. Our thanks for their creativity and dedication. Lesley Richardson contributed to the design of the studies, and the development and coordination of the data collection methods.

References

Banerjee S, Ramachandran G, Vadali M *et al.* (2014) Bayesian hierarchical framework for occupational hygiene decision making. *Ann Occup Hyg*; **58**: 1079–93.

Brooks SP, Gelman A. (1998) General methods for monitoring convergence of iterative simulations. *J Comp Graph Stat*; **7**: 434–55.

Browne WJ, Goldstein H, Rasbash J. (2001) Multiple membership multiple classification (MMMC) models. *Stat Model*; **1**: 103–24.

Carlin BP, Louis TA. (2009) *Bayesian methods for data analysis*. 3rd edn. Boca Raton, FL: CRC Press.

Department of Employment and Immigration. (1971) *Canadian classification dictionary of occupations*. **Vol. 1**. Ottawa, ON: Department of Employment and Immigration.

Doubleday A, Baker MG, Lavoué J *et al.* (2019) Estimating the population prevalence of traditional and novel occupational exposures in Federal Region X. *Am J Ind Med*; **62**: 111–22.

Friesen MC, Coble JB, Lu W *et al.* (2012) Combining a job-exposure matrix with exposure measurements to assess occupational exposure to benzene in a population cohort in Shanghai, China. *Ann Occup Hyg*; **56**: 80–91.

Friesen MC, Macnab YC, Marion SA *et al.* (2006) Mixed models and empirical bayes estimation for retrospective exposure assessment of dust exposures in Canadian sawmills. *Ann Occup Hyg*; **50**: 281–8.

Gelman A, Carlin JB, Stern HS *et al.* (2014) *Bayesian data analysis*. 3rd edn. Boca Raton, FL: CRC Press.

Gelman A, Hill J. (2007) *Data analysis using regression and multilevel hierarchical models*. New York, NY: Cambridge University Press.

Gelman A, Shirley K. (2011) Inference from simulations and monitoring convergence. In Brooks S, Gelman A, Jones GL *et al.*, editors. *Handbook of Markov chain Monte Carlo*. Boca Raton, FL: CRC Press.

Gérin M, Siemiatycki J, Kemper H *et al.* (1985) Obtaining occupational exposure histories in epidemiologic case-control studies. *J Occup Med*; **27**: 420–6.

Greenland S. (2000) Principles of multilevel modelling. *Int J Epidemiol*; **29**: 158–67.

Hall DB. (2000) Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*; **56**: 1030–9.

Jones RM, Burstyn I. (2016) Cross-classified occupational exposure data. *J Occup Environ Hyg*; **13**: 668–74.

Kruschke JK. (2015) *Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan*. 2nd edn. Boston, MA: Academic Press.

Labrèche F, Goldberg MS, Valois MF *et al.* (2010) Postmenopausal breast cancer and occupational exposures. *Occup Environ Med*; **67**: 263–9.

Lacourt A, Cardis E, Pintos J *et al.* (2013) INTEROCC case–control study: lack of association between glioma tumors and occupational exposure to selected combustion products, dusts and other chemical agents. *BMC Public Health*; **13**: 340.

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

Momoli F, Abrahamowicz M, Parent ME *et al.* (2010) Analysis of multiple exposures: an empirical comparison of results from conventional and semi-bayes modeling strategies. *Epidemiology*; **21**: 144–51.

Peters S, Vermeulen R, Portengen L *et al.* (2011) Modelling of occupational respirable crystalline silica exposure for quantitative exposure assessment in community-based case-control studies. *J Environ Monit*; **13**: 3262–8.

Peters S, Vermeulen R, Portengen L *et al.* (2016) SYN-JEM: a quantitative job-exposure matrix for five lung carcinogens. *Ann Occup Hyg*; **60**: 795–811.

Pintos J, Parent ME, Richardson L *et al.* (2012) Occupational exposure to diesel engine emissions and risk of lung cancer: evidence from two case-control studies in Montreal, Canada. *Occup Environ Med*; **69**: 787–92.

Plummer M. (2003) JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing. Vienna, Austria: Technische Universität Wien.

Portengen L, Linet MS, Li GL *et al.*; Chinese Center for Disease Control and Prevention—U.S. National Cancer Institute Benzene Study Group. (2016) Retrospective benzene exposure assessment for a multi-center case-cohort study of benzene-exposed workers in China. *J Expo Sci Environ Epidemiol*; **26**: 334–40.

Quick H, Huynh T, Ramachandran G. (2017) A method for constructing informative priors for Bayesian modeling of occupational hygiene data. *Ann Work Expo Health*; **61**: 67–75.

Ramachandran G. (2019) Progress in Bayesian statistical applications in exposure assessment. *Ann Work Expo Health*; **63**: 259–62.

Roberts B, Cheng W, Mukherjee B *et al*. (2018) Imputation of missing values in a large job exposure matrix using hierarchical information. *J Expo Sci Environ Epidemiol*; **28**: 615–48.

Sauvé JF, Siemiatycki J, Labrèche F *et al*. (2018) Development of and selected performance characteristics of CANJEM, a general population job-exposure matrix based on past expert assessments of exposure. *Ann Work Expo Health*; **62**: 783–95.

Siemiatycki J, Lavoué J. (2018) Availability of a new job-exposure matrix (CANJEM) for epidemiologic and occupational medicine purposes. *J Occup Environ Med*; **60**: e324–8.

Siemiatycki J, Wacholder S, Richardson L *et al*. (1987) Discovering carcinogens in the occupational environment. Methods of data collection and analysis of a large case-referent monitoring system. *Scand J Work Environ Health*; **13**: 486–92.

Toti S, Biggeri A, Baldasseroni A. (2006) A hierarchical Bayesian model for a variability analysis of measurements of occupational n-hexane exposure in Italy. *Stat Model*; **6**: 175–85.

Zeng F, Lerro C, Lavoué J *et al*. (2017) Occupational exposure to pesticides and other biocides and risk of thyroid cancer. *Occup Environ Med*; **74**: 502–10.