Ann. Occup. Hyg., 2016, Vol. 60, No. 7, 885–899
doi:10.1093/annhyg/mew029
Advance Access publication 1 June 2016

BOHS
The Chartered Society for Worker Health Protection

Evaluation of Automatically Assigned Job-Specific Interview Modules

Melissa C. Friesen[1*], and Qing Lan[1†], Calvin Ge[2], Sarah J. Locke[1], Dean Hosgood[3], Lin Fritschi[4], Troy Sadkowsky[5], Yu-Cheng Chen[1,6], Hu Wei[1], Jun Xu[1,7], Tai Hing Lam[7], Yok Lam Kwong[8,9], Kexin Chen[10], Caigang Xu[11], Yu-Chieh Su[12,13], Brian C. H. Chiu[14], Kai Ming Dennis Ip[7], Mark P. Purdue[1], Bryan A. Bassig[1], Nat Rothman[1‡] and Roel Vermeulen[2]

1.Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 9609 Medical Center Drive, North Bethesda, MD 20980, USA;
2.University of Utrecht, Utrecht, The Netherlands;
3.Department of Epidemiology & Population Health, Albert Einstein College of Medicine, Bronx, NY, USA;
4.School of Public Health, Curtin University, Perth, Australia;
5.Data Scientists Pty Ltd, Sunshine Coast, Queensland, Australia;
6.Environmental Health Research Center, National Health Research Institutes, Zhunan, Taiwan;
7.Division of Community Medicine and Public Health Practice, School of Public Health, The University of Hong Kong, Hong Kong;
8.Bone Marrow Transplant Unit, Queen Mary Hospital, Hong Kong;
9.Division of Haematology, Oncology and Bone Marrow Transplantation, Department of Medicine, The University of Hong Kong, Hong Kong;
10.Department of Epidemiology and Biostatistics, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China;
11.Department of Hematology, Hematology Research Laboratory and Pathology, West China Hospital of Sichuan University, Chengdu, Sichuan, China;
12.Division of Hematology-Oncology, Department of Internal Medicine, Buddhist Dalin Tzu Chi General Hospital, Chiayi, Taiwan;
13.School of Medicine, Tzu Chi University, Hualian, Taiwan;
14.Department of Public Health Sciences, University of Chicago, Chicago, IL, USA
*Author to whom correspondence should be addressed. Tel: +240-276-7278; fax: +240- 276-7835; e-mail: friesenmc@mail.nih.gov
†Co-first author.
‡Co-senior author.
Submitted 18 November 2015; revised 27 April 2016; revised version accepted 29 April 2016.

ABSTRACT

Objective: In community-based epidemiological studies, job- and industry-specific 'modules' are often used to systematically obtain details about the subject's work tasks. The module assignment is often made by the interviewer, who may have insufficient occupational hygiene knowledge to assign the correct module. We evaluated, in the context of a case–control study of lymphoid neoplasms in Asia ('AsiaLymph'), the performance of an algorithm that provided automatic, real-time module assignment during a computer-assisted personal interview.

Methods: AsiaLymph's occupational component began with a lifetime occupational history questionnaire with free-text responses and three solvent exposure screening questions. To assign each job to one of 23 study-specific modules, an algorithm automatically searched the free-text responses to the questions 'job title' and 'product made or services provided by employer' using a list of module-specific keywords, comprising over 5800 keywords in English, Traditional and Simplified Chinese. Hierarchical decision rules were used when the keyword match triggered multiple modules. If no keyword match was identified, a generic solvent module was assigned if the subject responded 'yes' to any of the three solvent screening questions. If these question responses were all 'no', a work location module was assigned, which redirected the subject to the farming, teaching, health professional, solvent, or industry solvent modules or ended the questions for that job, depending on the location response. We conducted a reliability assessment that compared the algorithm-assigned modules to consensus module assignments made by two industrial hygienists for a subset of 1251 (of 11 409) jobs selected using a stratified random selection procedure using module-specific strata. Discordant assignments between the algorithm and consensus assignments (483 jobs) were qualitatively reviewed by the hygienists to evaluate the potential information lost from missed questions with using the algorithm-assigned module (none, low, medium, high).
Results: The most frequently assigned modules were the work location (33%), solvent (20%), farming and food industry (19%), and dry cleaning and textile industry (6.4%) modules. In the reliability subset, the algorithm assignment had an exact match to the expert consensus-assigned module for 722 (57.7%) of the 1251 jobs. Overall, adjusted for the proportion of jobs in each stratum, we estimated that 86% of the algorithm-assigned modules would result in no information loss, 2% would have low information loss, and 12% would have medium to high information loss. Medium to high information loss occurred for <10% of the jobs assigned the generic solvent module and for 21, 32, and 31% of the jobs assigned the work location module with location responses of 'someplace else', 'factory', and 'don't know', respectively. Other work location responses had ≤8% with medium to high information loss because of redirections to other modules. Medium to high information loss occurred more frequently when a job description matched with multiple keywords pointing to different modules (29–69%, depending on the triggered assignment rule).
Conclusions: These evaluations demonstrated that automatically assigned modules can reliably reproduce an expert's module assignment without the direct involvement of an industrial hygienist or interviewer. The feasibility of adapting this framework to other studies will be language- and exposure-specific.

KEYWORDS: case-control studies; epidemiologic studies; occupational exposure; solvents

INTRODUCTION

Case–control studies that aim to evaluate health effects related to occupational risk factors typically collect occupational information using a lifetime occupational history questionnaire and, in many studies, supplemental occupation- and industry-specific job 'modules' that ask additional exposure-oriented questions. For example, a participant who reported having worked as a textile worker would be asked additional questions related to the textile industry using a textile module and a welder would be asked additional welding-related questions using a welder module. Use of these modules can reduce exposure misclassification by capturing important within-job exposure differences that occur both between- and within-subjects across time that would not be captured using occupation alone (Gerin *et al.*, 1985; Stewart *et al.*, 1998). The module responses can then be used to develop exposure decision rules to efficiently and transparently obtain exposure estimates for the study participants (Fritschi *et al.*, 2009; Behrens *et al.*, 2012; Pronk *et al.*, 2012; Friesen *et al.*, 2013; Carey *et al.*, 2014; Peters *et al.*, 2014).

The assignment of the appropriate module can be challenging during interviews because many interviewers have insufficient occupational hygiene knowledge with which to choose the appropriate module. As a result, use of these modules may be a two-step process: first, the participant is interviewed to obtain general occupational information, such as job title and task; second, after an occupational hygienist reviews the first interview's occupation information, the participant is re-interviewed with expert-assigned modules targeted to the reported jobs (e.g. Gerin *et al.*, 1985; Stewart et al., 1996, 1998; Fritschi *et al.*, 2005; MacFarlane

et al., 2012). To remove the burden of re-contacting and re-interviewing the subject, some studies have provided training to the interviewer to select the most appropriate module based on the occupational history information (e.g. Carey *et al.*, 2014). In some cases, the interviewer's selection has been aided by searching the responses entered during a computer-assisted interview with expert-derived keyword lists based on job title and task to provide the interviewer with a real-time, short list of modules from which to choose (e.g. Colt *et al.*, 2011). However, the error rate from interviewer module assignment is unknown and the training required to select the modules may be impractical in studies involving multiple sites and using a large team of interviewers.

To facilitate module assignment during a computer-assisted personal interview, we developed a computerized algorithm—NCI OccMATES: Occupational Modules Automatically Triggered in Epidemiologic Studies—that used free-text questionnaire responses to provide an automated, real-time assignment of each job to one of 23 modules. NCI OccMATES was implemented on study tablet computers to search the free-text entry of responses to a lifetime occupational history questionnaire against extensive lists of over 5800 module-specific keywords to identify keyword matches. Based on the keyword match(es), a single module was assigned for each job using a set of expert-derived hierarchical decision rules. This module was incorporated into the interviews immediately following the lifetime occupational history questions using OccIDEAS, a software application that provided the framework for storage and delivery of our exposure-oriented modules (Fritschi *et al.*, 2009). This occupational data collection and algorithm-module assignment were conducted within the Multi-Center Study of Lymphoid Neoplasms in Asia (hereafter, 'AsiaLymph'), a hospital-based case–control study that enrolled cases and controls in four study centers: Hong Kong, Chengdu, and Tianjin, China, and Kaohsiung, Taiwan. In this article, we describe the algorithm and the results of a reliability assessment that compared the algorithm-assigned modules to those assigned by two industrial hygienists using English translations of participants' responses to the occupational history interview.

METHODS

Occupational questionnaires

The occupational questionnaires used in AsiaLymph comprised a lifetime occupational history questionnaire and 23 modules with exposure-oriented questions focused on potential solvent exposure, including benzene, trichloroethylene, and formaldehyde. All questions were developed in English (M.C.F., S.J.L., R.V.) and translated into Traditional and Simplified Chinese (Y.C.C., H.W., J.X.). The translations of all occupational questions were reviewed by industrial hygienists from each study center to ensure location-specific nomenclature was incorporated. English-language modules are provided (Supplementary material is available at *Annals of Occupational Hygiene* online); Chinese language versions are available from the corresponding authors.

The occupational history questionnaire comprised, for each job reported by the subject, open-ended questions on 'what was the name of the employer or workplace', 'what was your job title', 'what did the employer make, or what service did they provide', and 'what were your main activities or duties', as well as questions on job start and stop years, and days per week and months per year worked in each job. In addition, three solvent exposure screening questions were asked:

- In this job, did you ever use paints, stains or varnishes or work in an area where they were used?
- In this job, did you ever use solvents, glues, degreasing agents (to clean metal parts), gasoline or other fuels, or work in an area where they were used?
- In this job, did you ever use particle board, plywood, or veneered woods or work in an area where they were used?

The exposure-oriented modules comprised 20 modules focused on specific occupations and industries (e.g. chemist module, healthcare module), two solvent modules that captured information on solvent-related tasks (solvent module, industry solvent module), and one work location module (Table 1). The modules were adapted (by M.C.F., S.J.L., R.V.) from previously used modules in NCI case–control studies and

Table 1. List of study modules, including an overview of within-module redirections to other modules and cross-module questions. See Supplementary Table S1, available at *Annals of Occupational Hygiene* online for more detail on what subsets of responses result in these redirections and cross-module questions.

Module, abbreviation	Includes re-directions to[a]	Cross-module questions												Proportion of all jobs (%)
		Degrease	Paint	Strip paint	Glue	Particle board	Handle fuel	Hands	Stain	Job type	QC sample collection	Molding plastic	Pesticide use	
Occupation & Industry-specific modules														
Chemist, CH								x			x			0.9
Chemical industry, CHI		x	x	x	x	x	x	x		x	x			2.3
Dry cleaning & textile industries, DLI_TXI	LEI	x	x	x	x	x	x	x		x	x			6.4
Embalming, EM														<0.1
Foundry industry, FOI		x	x	x	x	x	x	x		x	x			2.7
Furniture industry, FUI		x	x	x	x	x	x	x	x			x		1.3
Farming and food industry, GF_FDI		x	x	x	x	x	x	x		x			x	18.8
Health professional, HP		x	x	x	x	x	x	x						2.6
Janitor, JA		x	x		x		x	x						2.2
Leather industry, LEI		x	x	x	x	x	x	x		x		x		0.5
Lumber industry, LUI	FUI	x	x	x	x	x	x	x	x					0.9
Oil refinery industry, ORI		x	x	x	x	x	x	x		x	x			0.9
Pesticide applicators, PE								x					x	<0.1
Photographic industry, PHI		x	x	x	x	x	x	x		x	x			0.2
Plastic industry, PLI		x	x	x	x	x	x	x		x	x	x		1.2
Pulp and paper industry, PPI		x	x	x	x	x	x	x		x	x			0.4

Table 1. *Continued*

Module, abbreviation	Includes re-directions to[a]	Cross-module questions												Proportion of all jobs (%)
		Degrease	Paint	Strip paint	Glue	Particle board	Handle fuel	Hands	Stain	Job type	QC sample collection	Molding plastic	Pesticide use	
Printing industry, PRI		x	x	x	x	x	x	x		x	x			1.0
Rubber industry, RUI		x	x	x	x	x	x	x		x	x			0.2
Shoe industry, SHI		x	x	x	x	x	x	x		x				0.8
Teaching, TE		x	x					x						3.4
Generic modules														
Solvent, SOL	GF[b]	x	x	x	x	x	x	x	x					21.5
Industry solvent, INDSOL	GF[b]	x	x	x	x	x	x	x		x	x			0.3
Work location, BUP	GF[b], HP, INDSOL, SOL, TE													32.7

[a]Responses to certain questions redirected that job to questions in other modules. For example textile workers working with leather were asked questions from the leather industry module.
[b]Gardening and farming-related questions from the GF_FDI module.

incorporated additional knowledge obtained from previous studies conducted in China.

The solvent and industry solvent modules asked solvent task-related questions on degreasing, painting, paint stripping, gluing, fueling, hand contact with solvents, and working with particle board. The solvent module asked all of the solvent task-related questions, whereas in the industry solvent module the tasks that were queried depended on the occupation (e.g. administrative, production, management, quality control and engineering, maintenance, and material handling). For example, if the occupation was 'administrative or management', no solvent task questions were asked; if the occupation was 'material handling', only the paint, glue, board, and fuel questions were asked; and if the occupation was 'quality control, engineers, or other technical positions', questions on the collection and testing of production line samples were asked. These solvent task-related questions were also asked within many of the occupation- and industry-specific modules, which are identified as cross-module questions in Table 1 (for more detail see Supplementary Table S1, available at *Annals of Occupational Hygiene* online). In particular, these cross-module solvent task-related questions were triggered when the subject indicated that they did maintenance or utility work or were involved in shipping, receiving, or storage work.

The work location module was assigned whenever a job was not assigned to any of the occupation-, industry-, or solvent-specific modules, based on the algorithm described in the next section, to redirect jobs in specific work locations to appropriate modules. The module stated: 'The computer has not been able to accurately assess what kind of job you reported. We would like to verify if the current job is mostly: [participant asked to identify the best fit from 8 categorical work locations]'. Jobs with work locations of farm, hospital, school, factory, or construction site work locations were redirected to farming, health professional, teaching, industry solvent, or solvent modules, respectively. If the work was in any other location (e.g. office, store, restaurant, someplace else) no more questions were asked for that job.

Keyword development

We developed lists of over 5800 occupation and industry keywords in English (n = 1580), traditional Chinese (Hong Kong, Kaohsiung, n = 2422) and simplified Chinese (Chengdu, Tianjin, n = 1892) that were specific to one of the 23 modules. One set of occupational keywords was developed to search the subject's responses to the questions 'job title' to identify keywords linked to occupations associated with each module (21 occupation sets). For example 'seamstress' was a keyword linked to dry cleaning and textile industry occupations. Another set of industrial keywords was developed to search the responses to 'product made or services provided by employer' to identify keywords linked to industries associated with each module (17 industry sets). For example 'smelting' was a keyword linked to the foundry industry.

All keyword sets were first developed in English (M.C.F., S.J.L., R.V.). This team supervised the translation of the English-language lists by three native Chinese speakers (Y.C.C., H.W., J.X.) into the local languages of the four study centers (simplified Chinese for Chengdu and Tianjin; traditional Chinese for Kaohsiung and Hong Kong). The translation process included generation of additional, similar local words and phrases that may be used to describe job title or employer activity. The translations were then reviewed by occupational hygienists from each of the four study centers, with additional words and revisions incorporated by the development team.

Each keyword set contained three types of keywords that were used alone because the word or word string uniquely identified relevant occupation or industry (Type 1) or were used in combination with another keyword to identify the relevant occupation or industry (Types 2 and 3). Examples of Type 1 keywords were 'teacher' and 'dry cleaner', which were considered sufficient information with which to assign the 'Teacher' module and 'Dry Cleaning and Textile Industry' module, respectively. Type 2 and 3 keywords were designed because not all processes and activities could be described succinctly or completely by specific words or word strings or because the word order may vary. In general, Type 2 keywords described the material or service being processed or provided (e.g. clothes, dry-cleaning), and Type 3 keywords described the process, place or person related to the product or service (e.g. bleaching, workshop, worker). A Type 2 or Type 3 keyword could each appear in multiple keyword sets to capture similar work activities (e.g. Type 3 keyword 'worker'). However, unique combinations of Type 2 and Type 3 keywords were

designed to occur only within one keyword set. To match a keyword set, the response had to include a Type 1 match or both a Type 2 and a Type 3 match, in any order (hereafter, Type 2/3 match). For example, the responses 'bleaching clothes' and 'dry-cleaning worker' would both have a positive match to the occupation keyword set for the 'Dry Cleaning and Textile Industry' module.

Algorithm description

The occupational and industry keyword sets were used to search the occupational history responses to assign the most appropriate module. All jobs received a module. The occupational keyword sets were used to search the job title responses from the occupational history; this search either identified no matches to any keyword set, single match (Type 1 or Type 2/3), or multiple matches (of any match type). Similarly, the industrial keywords were used to search the employer activity responses to identify no match, single match, or multiple industrial keyword set matches.

The varying combinations of match results were processed using a set of hierarchical decision rules that assigned a single module for each reported job according to the flow chart shown in Figure 1. See Supplementary Table S2, available at *Annals of Occupational Hygiene* online for the individual rules and actions. If no module-specific keyword was identified for a job, the module was assigned based on the solvent screening questions: if 'yes' to any of these questions, the solvent module was assigned (rule #1); and if 'no' to all these questions,

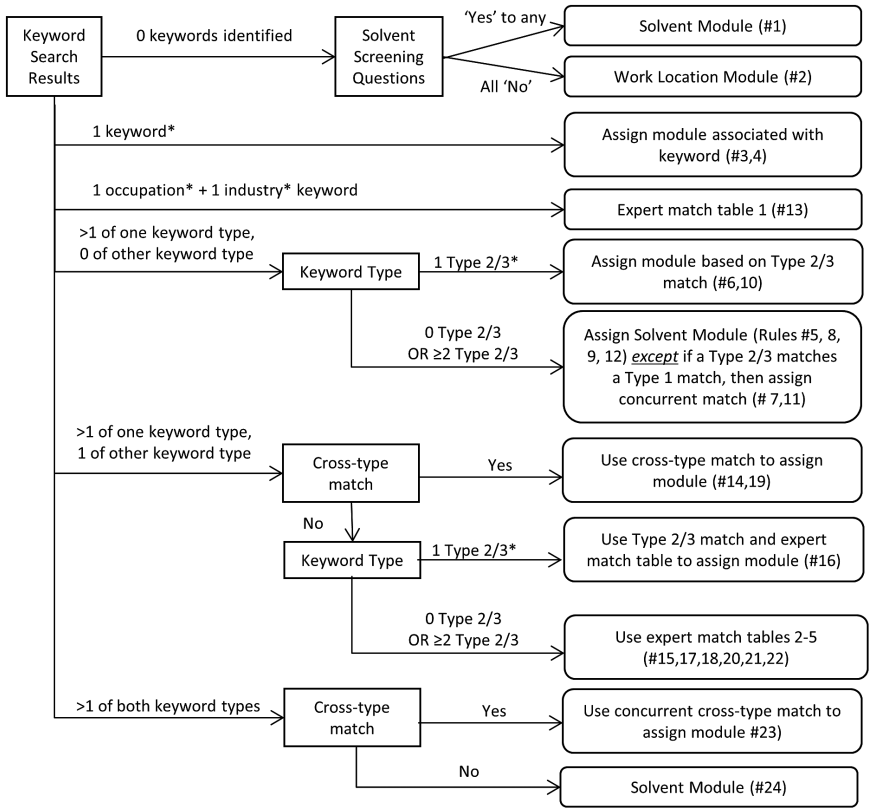

Figure 1 Hierarchical algorithm decision rules to assign modules. The rules were based on the number of keyword sets identified in the response (occupation and/or industry), the keyword type (Type 1, Type 2/3), and whether the occupation and industry keywords assigned the same or different modules ('cross-type match' = yes or no, respectively). Flow chart steps for single sets or single keyword types marked with * also include the rare scenario where multiple matches were found but all matches were within the same occupation or industry keyword set.

the work location module was assigned (rule #2). If the keyword(s) matched a single module based on the occupation *or* industry keyword sets, the corresponding module was assigned (single industry module: rule #3; single occupation module: rule #4).

If multiple keywords were identified for a job, the module was assigned using expert-derived rules based on the number of keyword sets identified in the response (occupation and/or industry), the keyword type (Type 1, Type 2/3), and whether the occupation and industry keywords assigned the same or different modules ('cross-type match' = yes or no, respectively). These rules fall into three distinct groups: multiple occupation keywords but no industry keywords identified (rules #5–8), multiple industry keywords but no occupation keywords identified (rules #9–12), or combinations of both occupation and industry keywords identified (rules #13–24). Within each group, the rules designated what to do when matches to multiple occupation and/or industry sets occurred. When matches to multiple occupation or industry keyword sets occurred, the rules prioritized the cross-type module match (rules #14, 19, 23). For example, if multiple occupation keyword matches were found, the rules prioritized the match that assigned the same module as the industry keyword match. Otherwise, the rules prioritized Type 2/3 matches, which used detailed adjectives and were not order specific, over Type 1 matches (rules #6, 7, 10, 11, 16). When the occupation and industry matches directed to different modules (cross-type match = no), the module assignment was made using expert-derived match tables (rules #13, 15–23). In the expert match tables, the industry solvent module was assigned when the participant reported working in a plant or factory and did not identify the type of factory, or when the keywords identified multiple industries. When the match combination was deemed too ambiguous, the generic solvent module was assigned (rule #24 and within expert match tables).

Study implementation of questionnaires and algorithm

The occupational history component was administered during comprehensive interviews conducted by trained interviewers using tablet computers with study-specific software. The interviewers were trained to prompt the participants for detailed responses, which were entered into the software. The software used the algorithm to automatically search the 'job title' and 'product made or services provided by employer' responses to assign one of the 23 modules to each job reported by the participant. The study-specific software launched the assigned modules for that participant immediately after the occupational history section, using a stand-alone version of OccIDEAS that incorporated the study-specific modules (Fritschi *et al.*, 2009).

Job selection for reliability assessment

Overall, 11 409 jobs were reported between 1 September 2013, when final versions of the modules and algorithm were in place and interviewer training was completed, and 31 January 2015, when the evaluations reported here began. To compare the algorithm-assigned module to the module that an industrial hygienist would have assigned in a reliability assessment, we selected a subset of these jobs using a stratified randomized selection (without replacement) procedure using strata defined by the algorithm-assigned module. Most modules were infrequently assigned, except the solvent and work location modules (Table 1). For each module except the solvent and work location modules, we selected all jobs if the module was assigned to ≤25 jobs and randomly selected 25 jobs if the module was assigned to 26–249 jobs, 50 jobs if the module was assigned to 250–499 jobs, and 100 jobs if the module was assigned to ≥500 jobs. For the solvent module, we randomly selected 100 jobs that were assigned the solvent module based on keyword-module linkage (55% of solvent module assignments, rules #3–24) and 100 jobs that were assigned that module because no keywords were identified but the participant responded 'yes' to any of the three solvent screening questions (44% of all solvent module assignments, rule #1). For the work location module, job selection was stratified by work location responses. We selected all jobs in the location category if the location category was reported for ≤25 jobs and randomly selected 25 jobs if the category was reported for 26–249 jobs, 50 jobs if the category was reported for 250–499 jobs, and 100 jobs if the

category was reported for ≥500 jobs. In total, 1251 jobs were selected.

Expert module assignment

Two industrial hygienists (S.J.L., C.G.) independently reviewed English translations of the 1251 occupational history questionnaire responses in the reliability subset and assigned the most appropriate module to each job. S.J.L. had been involved in the original keyword and algorithm development and C.G. had no prior involvement. The experts were blind to the algorithm module assignments and to the module question responses. Jobs in which the two experts disagreed on the module assignment were re-reviewed by the same two experts to obtain a consensus assignment. Jobs where the assignments still differed were reviewed by M.C.F. (blind to the algorithm module assignment), who had been involved in all aspects of the keyword and algorithm development.

Expert review of assigned module coverage of pertinent questions

The two experts (S.J.L., C.G.) were provided with the subset of jobs where the expert consensus and algorithm assignments differed (n = 483). This subset excluded jobs where the discordance was between the solvent and the industry solvent module because these modules were nearly identical (n = 46). The experts independently reviewed the job description, the expert consensus module assignment, the algorithm module assignment, and the work location category for those assigned the work location module. Each expert considered whether the assigned module included the questions that were likely to be most relevant to that job (e.g. did they paint?). Relevant questions that were missed were considered to be a potential 'information loss' for the exposure assessment process. Each expert provided a qualitative estimate of the degree of potential exposure information lost by the assigned module from the missed questions for each job, using four subjective categories:

1. *All* pertinent questions covered. No loss of information.
2. *Most* pertinent questions covered. Low loss of information.
3. *Some* pertinent questions covered. Medium loss of information.
4. *Few* (or no) pertinent questions covered. High loss of information.

Potential information loss was more likely to occur when pertinent industry-specific questions were missed because the cross-module questions were included in most modules. For example, if the most relevant questions for a job were about painting and gluing and the assigned module included those questions, the expert would have provided a rating of '1', no loss of information. In a second example, if the expert consensus module was the furniture module and the assigned module was the solvent module, the expert would have provided a rating of '1' (no information loss) if the expert considered the most relevant questions were covered by the cross-module questions, or a rating of '3' (medium information loss) if the most relevant questions were about the participant's role in manufacturing furniture. The experts were asked to not consider the expected answer (e.g. yes or no to the paint question) in the evaluation of information loss. The impact of the information loss on exposure decisions will be evaluated in future analyses. The two experts' degree of loss ratings had very good agreement (% agreement = 76%; kappa = 0.61; weighted kappa = 0.84); thus, consensus 'degree of loss' ratings were not obtained.

Statistical analyses

We calculated the proportion of jobs in the reliability subset where the two experts agreed on the module assignments. Similarly, we calculated the proportion of jobs where the expert consensus and algorithm module assignments agreed, as raw agreement in the reliability subset and as estimated agreement extrapolated to all 11 409 jobs. The extrapolated agreement was obtained by first calculating strata-specific agreement and then weighting that agreement by the proportion of jobs in those strata. For example the strata-specific agreement for the chemist module, farming and food industry module, and industry solvent module contributed to 0.9, 18.8, and 0.3% of the extrapolated estimate, respectively (Table 1). Evaluations of the proportion of jobs with varying degrees of information loss were based on the average of the two experts' ratings, categorized as follows:

1 = both experts indicated no loss; >1–2 = low loss; >2–3 = medium loss; and >3 = high loss.

RESULTS

Keyword matches

In the 11 409 jobs reported during the sample period, an average of 1.4 keyword matches per job was identified by the algorithm. The three most frequently identified industry keywords were Chinese translations of 'cultivation' (in 979 jobs), 'grain' (in 447 jobs), and 'cultivating' (in 446 jobs). The three most frequently identified occupation keywords were 'farmer' (in 750 jobs), 'field hand' (in 204 jobs), and 'repair' (in 185 jobs).

Expert agreement

The proportion of the 11 409 jobs that the algorithm assigned to each module is provided in Table 1. The most frequently assigned modules were the work location (32.7%), solvent (21.5%), farming and food industry (18.8%), and dry cleaning and textile industries (6.4%) modules. Other modules accounted for <0.1 to 3.4% of all jobs.

The two experts' independent module assignments matched for 80.3% of the 1251 jobs. Discordance between expert assignments to the solvent and industry solvent modules accounted for an additional 6.2% of the jobs. The remaining 246 jobs were re-reviewed to derive a consensus assignment; 87 of these jobs required review by a third expert.

Algorithm versus expert agreement and information loss: Overall

In comparison to the expert consensus-assigned modules, the algorithm assignment had exact matches for 722 (57.7%) of the 1251 jobs in the reliability assessment (Table 2). An additional 46 (3.7%) jobs differed because one approach assigned the solvent module and the other assigned the industry solvent module. The remaining 483 jobs (38.6%) represented disagreements between the consensus and algorithm assignments and were reviewed to identify the potential information lost.

Evaluations based on potential information loss showed that the algorithm's discordant assignment resulted in no information loss for the majority of these jobs (51.1% of discordant jobs, 19.7% of jobs in reliability subset; Table 2). Low, medium, or high information losses were estimated to occur for 2.7, 6.4, and 9.8% of the jobs in the reliability assessment, respectively. Because we oversampled infrequent module assignments, extrapolation to all jobs resulted in a higher estimated proportion of exact matches (67.5% of all jobs) and lower proportions of medium and high information losses (5.3 and 6.2%, respectively) than in the reliability

Table 2. Overall: agreement between algorithm- and expert-assigned modules, with an assessment of the degree of information loss when the two approaches were discordant

	Jobs in reliability study (n = 1251)		Extrapolated proportion (%) of all jobs[b]
	N	%[a]	
Exact match	722	57.7	67.5
No match			
Solvent/industry solvent mismatch	46	3.7	3.8
No information loss	247	19.7	15.0
Low information loss	34	2.7	2.3
Medium information loss	80	6.4	5.3
High information loss	122	9.8	6.2

[a]Proportions are unadjusted for sampling weights.
[b]Extrapolation to all jobs was calculated by weighting each stratum-specific agreement by the proportion of all jobs observed in that stratum. For example the strata-specific agreement for the Chemist Module, Farming and Food Industry Module, and Industry Solvent Module accounted for 0.9, 18.8, and 0.3% of the extrapolated estimate, respectively. Proportions of all jobs each stratum represented are shown in Table 1 for most modules and Table 4 for the Solvent and Work Location Modules.

subset. Overall, an estimated 86.3% of the algorithm's assignments were consistent with the consensus assignment or would have no information loss (67.5 + 3.8 + 15.0). An additional 2.3% would be expected to have low information loss and 11.5% would have potentially medium to high information loss.

Algorithm versus expert agreement and information loss: By algorithm rule and module

Evaluations stratified based on groups of algorithm rules that applied to each assignment are shown in Table 3. Medium or high potential information losses were ≤15% when no keywords were identified (rule#1, #2) and when only a single occupation and/or industry keyword was identified (rule #3, #4, #13). These scenarios occurred for 94% of all jobs. The algorithm identified multiple matches within a keyword type for the remaining 6% of jobs. For these multiple matches, medium or high information losses ranged from 24 to 65% (rules #5–12, 14–24).

Evaluations conducted within the solvent, industry solvent, and work location modules are shown in Table 4. For jobs assigned the solvent module (20.5% of all jobs), medium or high potential information losses were observed for 7% of the jobs when the module was triggered by the exposure screening questions and 10% when triggered by algorithm rules. The industry solvent module was assigned only 0.3% of the time, but its assignment was estimated to have medium or high information loss for 48% of the jobs. For jobs receiving the work location module (33% of all jobs), medium or high information losses were less than 8% for most work location categories, with three exceptions. A higher proportion with medium or high information loss was observed for locations of 'someplace else' (21%), 'factory/warehouse' (32%), and 'don't know/missing' (31%).

The two most frequently assigned industry-specific modules—farming and food industry and dry cleaning and textile industries modules—had low, medium, or high information loss in less than 2% of the jobs assigned each module (not shown). Module-specific evaluations were not reported for the other modules because of their low prevalence.

DISCUSSION

This study demonstrated that an expert-designed automated algorithm provided real-time module assignments during computer-assisted personal interviews that reliably reproduced post hoc module assignments made by industrial hygienists in this study. Overall, we estimated that 86% of all algorithm module assignments would result in no potential exposure information loss, 2% would have low information loss, and 12% would have medium or high information loss. Evaluations in strata based on groups of algorithm rules and the assigned module provided important insights into directions for future improvements of the algorithm. To our knowledge, no similar evaluations comparing interviewer or automated module assignments to those of occupational hygienists have been previously reported.

The two most prevalent modules—the solvent module (assigned to 20.5% of jobs) and the work location module (assigned to 33% of jobs)—generally had ≤10% of the jobs with medium to high information loss. This good performance occurred because the solvent module captured the majority of solvent-exposed tasks and the work location module redirected participants reporting work locations with potential solvent exposure (e.g. healthcare, farming) to the appropriate module. In the work location module, some refinements may reduce information loss for the 'factory' and 'someplace else' responses. Responses of 'factory' resulted in an estimated medium to high information loss for 32% of the jobs reporting that location because these responses were redirected to the industry solvent module, which does not include important industry-specific questions. This could be refined by asking an additional question about whether the factory was in any of the industries of interest for which modules were developed to redirect that job to the appropriate module. Responses of 'someplace else' resulted in medium to high information loss for 21% of the jobs reporting that location, which suggests that an important work location category may have been missed, but the closed-ended design of this response category did not allow us to explicitly evaluate which locations were missed. However, these modest potential information losses are likely overestimates, because to be assigned this module the participants had to respond 'no' to each of the three solvent screening questions and no keyword matches were identified from the occupational history responses.

Potential information loss was more prevalent when multiple keyword matches were identified in the occupational history (29–65% medium/high loss, depending on rule). Fortunately, this occurred in <6% of all

Table 3. By algorithm rule: agreement between algorithm- and expert-assigned modules, with an assessment of the degree of information loss when the two approaches were discordant.

Exposure screening question with yes response[a]	No. of occupation keywords identified	No. of industry keywords identified	Algorithm rule numbers	No. of jobs	% of all jobs	Proportion of jobs in strata (%)			
						Exact match	No info. loss or INDSOL/SOL discordance	Low, info. loss	Medium or high info. loss
Yes	0	0	1	100	9.0	58	30	5	7
No	0	0	2	311	32.5	48	32	5	15
—	0	1	3	229	12.4	54	30	1	15
—	1	0	4	265	20.5	67	18	1	14
—	>1	0	5–8	26	1.7	8	23	4	65
—	0	>1	9–12	15	0.9	40	20	0	40
—	1	1	13	241	19.8	73	10	2	15
—	1	>1	14–18	25	1.2	36	24	4	36
—	>1	1	19–22	34	1.4	50	21	6	24
—	>1	>1	23–24	4	0.4	50	0	25	25
—	Missing	Missing	Missing	1	0.3	100	0	0	0
Overall[c]		Reliability subset		1251		58	23	3	16
		All jobs (estimated)		11 409		68	18	2	12

[a]Screening questions were not used except in rules #1 and #2. No indicates that a 'no' response was received to all three exposure screening questions. Yes indicates that at least one screening question had a 'yes' response.
[b]See Supplementary Table S1, available at *Annals of Occupational Hygiene* online for more detail on each rule's criteria and resulting action.
[c]See Table 2. Provided here for comparison purposes.

Table 4. Generic solvent and work location modules: agreement between algorithm- and expert-assigned modules, with an assessment of the degree of information loss when the two approaches were discordant

Module	Sub-group	No. of jobs in reliability study	% of all jobs	Proportion of jobs in strata (%)			
				Exact match	INSOL/SOL discordance or no info. loss	Low, info. loss	Medium or high info. loss
Solvent (SOL), *reason for assignment*							
	Assigned based on screening questions, no keywords identified	100	9.0	58	30	5	7
	Assigned based on identified keywords	100	11.5	55	30	5	10
Industry solvent (INDSOL)		25	0.3	12	36	4	48
Work location (BUP), *response category*[a]							
	Factory/warehouse	25	3.4	28	28	12	32
	School	25	0.6	32	64	0	4
	Store/restaurant	25	4.9	48	44	0	8
	Office	25	13.9	92	0	0	8
	Construction	50	2.8	14	80	2	4
	Someplace else	100	6.6	70	2	7	21
	Farm	21	0.2	10	86	0	5
	Hospital	6	0.1	0	100	0	0
	DK/missing/skipped	35	0.3	60	0	9	31

[a]Participant was asked which category best described his or her work location. No additional information was collected on the work location if the participant responded 'Someplace else'.

jobs. The information loss generally occurred when appropriate industry-specific solvent exposure questions were not asked, whereas solvent task-related questions were generally captured within imperfect module assignments. For example the industry solvent module, with 48% of the jobs with medium/high loss, was assigned when there were multiple keyword matches pointing to different modules or when there was too little information to determine the type of industry. Module-specific evaluations also showed that some keywords thought to be specific to a module were likely not specific enough. For example translations of the keywords 'wood', 'saw', and 'planing' used for the lumber industry resulted in many other jobs working with wood being assigned that module, whereas the more relevant module might be the furniture industry for furniture workers and solvent module for construction or other trade laborers. In another example, translations of the words 'testing' and 'analysis' assigned many workers involved in quality control/quality assurance to the chemist module, whereas an industry-specific module may have been more appropriate. This latter example could be addressed with refinements to the hierarchical set of rules to prioritize the triggered industry rather than the chemist occupation.

Incorporating OccMATES to provide a real-time assignment during the personal interview had several strengths. Most importantly, the same

module assignment procedure was used for all participants, regardless of the study center and interviewer's occupational hygiene expertise; thus reducing interviewer module selection bias. It also reduced study and respondent burden by conducting both components of the occupational data collection concurrently. The assignment of the modules was also transparent and the approach can be updated based on these evaluations. In addition, exposure assessments for jobs assigned an imperfect module will still be performed, but will require expert review rather than the automated assignment of exposure assessment decision rules (Fritschi *et al.*, 2009). Our findings suggest that expert review in this study should be focused on jobs with multiple matches within a keyword type and those indicating 'factory' or 'someplace else' in the work location module.

These evaluations also have several limitations that should be considered in interpreting these findings. First, the expert consensus module assignments represented only an 'alloyed' gold standard. The expert assignments were made post-interview, using job descriptions translated from Chinese to English. Thus, they were made using a more limited context than if it would have taken place during the interview, although they were made with much greater occupational expertise than the interviewer. We observed a modest degree of variability between the two experts, with only 13% of the jobs requiring a consensus review, of which 30% required a third expert to provide the final assignment. Second, because we excluded interviewer involvement in the module assignment, we were unable to obtain a measure of the error rate that would occur if an interviewer selected a module from a short list for comparison purposes. Third, to provide a conservative test of the algorithm, we provided the experts access to responses to all occupational history responses, whereas the algorithm module assignment was made solely based on two of these questions (job title, products made/services provided). For example the experts could use important information about industry that was reported in the employer's name response (which was often descriptive). Future refinements of the algorithm may include keyword searches of responses to these additional questions. For instance, we could use the job task information reported to develop keyword lists for tasks associated with each module to improve the specificity of the assignments. Finally, our categorization of potential 'information loss' considers only whether questions of interest were asked and not the likely response to the question or whether the job was likely exposed to solvents. The impact of this information loss on exposure decisions will be evaluated in future application of exposure decision rules.

In summary, this computerized algorithm provided a real-time module assignment that reliably reproduced an expert's module assignment, with limited potential information loss. Our findings are study specific. However, the framework could be extended to other studies. The degree of adaptation required will depend on the languages and exposures of interest and may require substantial changes to the modules, keyword lists (including adding common misspellings), and algorithm decision rules.

SUPPLEMENTARY DATA

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

ACKNOWLEDGMENTS

This project was funded by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, NCI, NIH. LF is supported by a fellowship from the National Health and Medical Research Council of Australia. The authors report no conflicts of interest. The authors wish to acknowledge the study interviewers, local industrial hygienists, all other study staff, and the patients that participated in the study.

REFERENCES

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

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

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

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.

Fritschi L, Benke G, Hughes AM *et al.* (2005) Risk of non-Hodgkin lymphoma associated with occupational exposure to solvents, metals, organic dusts and PCBs (Australia). *Cancer Causes Control*; 16: 599–607.

Fritschi L, Friesen MC, Glass D, Benke G, Girschik J, Sadkowsky T. (2009) OccIDEAS: retrospective occupational exposure assessment in community-based studies made easier. *J Environ Public Health*; 2009: 957023.

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

MacFarlane E, Benke G, Sim MR *et al.* (2012) OccIDEAS: an innovative tool to assess past asbestos exposure in the Australian Mesothelioma Registry. *Saf Health Work*; 3: 71–6.

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-controls versus expert review of individual jobs: Programmable decision rules versus expert review of individual jobs. *Occup Environ Med*; 69: 752–8.

Stewart PA, Stewart WF, Heineman EF *et al.* (1996) A novel approach to data collection in a case-control study of cancer and occupational exposures. *Int J Epidemiol*; 25: 744–52.

Stewart PA, Stewart WF, Siemiatycki J *et al.* (1998) Questionnaires for collecting detailed occupational information for community-based case control studies. *Am Ind Hyg Assoc J*; 59: 39–44.