Annals of Work Exposures and Health, 2023, **67**, 663–672
https://doi.org/10.1093/annweh/wxad002
Advance access publication 3 February 2023
Original Article

The Chartered
Society for Worker
Health Protection

OXFORD

Automated Coding of Job Descriptions From a General Population Study: Overview of Existing Tools, Their Application and Comparison

Wenxin Wan[1], Calvin B. Ge[1], Melissa C. Friesen[2], Sarah J. Locke[2], Daniel E. Russ[2], Igor Burstyn[3], Christopher J. O. Baker[4], Anil Adisesh[5,6], Qing Lan[7], Nathaniel Rothman[2], Anke Huss[1], Martie van Tongeren[8], Roel Vermeulen[1] and Susan Peters[1,*]

[1]Department Population Health Sciences, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands;
[2]Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA;
[3]Department of Environmental and Occupational Health, Drexel University, Dornsife School of Public Health, Philadelphia, PA, USA;
[4]Department of Computer Science, Faculty of Science, Applied Science and Engineering, University of New Brunswick, Saint John, NB, Canada;
[5]Division of Occupational Medicine, Department of Medicine, University of Toronto, Toronto, ON, Canada;
[6]Department of Medicine, Dalhousie Medicine New Brunswick, Saint John, NB, Canada;
[7]Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Rockville, MD, USA;
[8]Centre for Occupational and Environmental Health, School of Health Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK

[*]Author to whom correspondence should be addressed. Tel: +31-30-253-7517; e-mail: s.peters@uu.nl

Abstract

Objectives: Automatic job coding tools were developed to reduce the laborious task of manually assigning job codes based on free-text job descriptions in census and survey data sources, including large occupational health studies. The objective of this study is to provide a case study of comparative performance of job coding and JEM (Job-Exposure Matrix)-assigned exposures agreement using existing coding tools.

Methods: We compared three automatic job coding tools [AUTONOC, CASCOT (Computer-Assisted Structured Coding Tool), and LabourR], which were selected based on availability, coding of English free-text into coding systems closely related to the 1988 version of the International Standard Classification of Occupations (ISCO-88), and capability to perform batch coding. We used manually coded job histories from the AsiaLymph case-control study that were translated into English prior to auto-coding to assess their performance. We applied two general population JEMs to assess agreement at exposure level. Percent agreement and PABAK (Prevalence-Adjusted Bias-Adjusted Kappa) were used to compare the agreement of results from manual coders and automatic coding tools.

Results: The coding per cent agreement among the three tools ranged from 17.7 to 26.0% for exact matches at the most detailed 4-digit ISCO-88 level. The agreement was better at a more general level of job coding (e.g. 43.8–58.1% in 1-digit ISCO-88), and in exposure assignments (median values of PABAK coefficient ranging 0.69–0.78 across 12 JEM-assigned exposures). Based on our testing data, CASCOT was found to outperform others in terms of better agreement in both job coding (26% 4-digit agreement) and exposure assignment (median kappa 0.61).

Conclusions: In this study, we observed that agreement on job coding was generally low for the three tools but noted a higher degree of agreement in assigned exposures. The results indicate the need for study-specific evaluations prior to their automatic use in general population studies, as well as improvements in the evaluated automatic coding tools.

Keywords: automatic job coding tool; free-text job description; general population studies; reliability

Received: September 22, 2022. Accepted: January 9, 2023.

What's Important About This Paper?

Applying automatic tools to code free-text job descriptions from large population studies could provide a cheaper and faster alternative than manual coding. This study compared the performance of available job coding tools with a single large-scale population data set and found generally low coding agreement. The results suggest the need for study-specific evaluations before their automatic use in population-based occupational studies, as well as improvements in the evaluated automatic coding tools.

Introduction

Standardized classification of occupations is an important step in occupational health research. In large general population studies, manual coding of free-text job descriptions can be both expensive and time-consuming. For a study with around 200 000 job entries, it would take >2 years for one coder to finish the work at a reasonable rate of coding (e.g. 1500 per week). Based on the recent inventory of occupational cohorts within the Network on the Coordination and Harmonisation of European Occupational Cohorts (OMEGA-NET) (Kogevinas *et al.*, 2020), about 40% of cohorts were estimated to have yet uncoded occupational data (personal communication M. Turner, December 2021), limiting the opportunities to advance occupational health research.

Several coding tools have been developed to automate or partially automate job coding in recent years (Burstyn *et al.*, 2014; Russ *et al.*, 2016; Warwick Institute for Employment, 2018; Kouretsis *et al.*, 2020). While some tools reported positive results [e.g. 80% match manual coding with high data quality and coding confidence (Warwick Institute for Employment, 2018)], no formal assessment has been conducted across existing automatic job coding tools using a single large-scale population study. For epidemiologic analyses, if the reliability of job coding methods and their impact on uncertainty in exposure assignment is evaluated, this information can be useful to adjust epidemiologic analyses for biases that these errors may create (Burstyn *et al.*, 2018).

We aim to provide an overview of existing automatic job coding tools, compare the agreement between coding tools and human reviewers along with inter-coder agreement, and the agreement in exposure assignments by job-exposure matrices (JEMs). The results inform researchers about the current development of automatic job coding tools and provide insights into the application of those tools in general population studies.

Methods

Automatic occupation coding tools

We inventoried existing job coding tools and their features (Table 1). We selected three coding tools [AUTONOC, CASCOT (Computer-Assisted Structured Coding Tool), and LabourR] to compare their coding performance. The inclusion was based on:

1). Availability, either publicly or via its developers;
2). Ability to assign job codes from free-text job descriptions in English;
3). Ability to code into systems closely related to the 1988 version of the International Standard Classification of Occupations (ISCO-88); and
4). Capability in batch coding (process multiple job entries at once).

AUTONOC is an online automatic job coding platform that facilitates automated occupational encoding. It uses the ENENOC (Ensemble Encoder for the National Occupational Classification) which is based on several machine learning techniques for sentence embedding acting as a hierarchical ensemble classifier algorithm (Garcia *et al.*, 2021). AUTONOC uses only job titles as input data and produces coding outputs in 2016 version of Canadian National Occupational Classification (NOC 2016), 2010 version of US Standard Occupational Classification (SOC 2010), and links to the Occupational Information Network (O*NET). Outputs in NOC 2016 were used for this analysis. ENENOC was benchmarked against a previous natural language processing (NLP) algorithm called ACA-NOC (Bao *et al.*, 2020) which was also developed for auto-coding to the Canadian National Occupational Classification. ENENOC showed improved performance on a curated test set (Garcia *et al.*, 2021) and was selected for use in this study.

CASCOT performs automatic job coding using free-text job titles to ISCO-08 and a number of UK job classification systems (Warwick Institute for Employment, 2018). Job coding is also possible with free-text job titles in several different languages (Table 1). CASCOT is built in the Java language and utilises a matching algorithm that matches not just index entries but examines the number and similarities of competing index entries. CASCOT produces "confidence scores" (ranging from 0 to 100), mimicking the Bayesian probability that the computer-assigned code is that which

Table 1. Overview of automatic job coding tools.

Automatic coding tool	Free-text input	Job coding output	Input language options	References
AUTONOC	Job title, industry	NOC 2016, SOC 2010, O-NET	English	Suarez Garcia *et al.* (2021)
CASCOT	Job title	UK SOC 1990, 2000, 2010; ISCO-08	Arabic; Chinese; Dutch; English; Finnish; French; German; Hindi; Indonesian; Italian; Portuguese; Romanian; Russian; Slovak; Spanish	Warwick Institute for Employment (2018)
LabourR	Job title	ISCO-08	English	Kouretsis *et al.* (2020)
NIOCCS	Job title, industry title	US SOC 2000, 2010	English	Russ *et al.* (2016)
SOCcer	Job title, job task, industry title	US SOC 2010	English	Russ *et al.* (2016)
SOCEye	Job title	US SOC 2010	English	Burstyn *et al.* (2014)
SOIC	Industry and occupations	BOC 1990	English	Patel *et al.* (2012)
O*Net	Job title, industry, task, etc.	O*NET- SOC	English	–
CAPS-Canada	Job title, industry, task, etc.	ISCO 1968, 1988, 2008; CCDO 1971-1989; NOC 2011; SOC 2010	English, French	–
CAPS-France	Job title, industry, task, etc.	ISCO 1968, 1988; PCS 1982, 2003	English, French	–
PROCODE	Job title	PCS 2003	English, German, French, Italian	Savic *et al.* (2022)

Note that O*Net, CAPS-Canada/France are search engines based on free-text inputs for individual job code searching.
NOC 2016, Canadian National Occupational Classification 2016 version; SOC, US Standardized Occupation Classification; SIC, Standardized Industry Classification; ISCO-88, 1988 version of the International Standard Classification of Occupations; ISCO-08, 2008 version of the International Standard Classification of Occupations; CCDO, Canadian Classification and Dictionary of Occupations; PCS, Procedure Coding System.

would be assigned manually by experts in job coding (Warwick Institute for Employment, 2018).

LabourR is a package in the R statistical software that was initially developed to retrieve the work experience history from a Curriculum Vitae (CV) to support data analysis from the Europass online CV editor (Kouretsis *et al.*, 2020). Based on the ESCO (European Skills, Competences, Qualifications and Occupations) hierarchical classification model, LabourR processes free-text job title input in English and provides coding outputs in ISCO-08. The tool is based on keyword look-up tables and a k-nearest neighbour classification algorithm (Kouretsis *et al.*, 2020).

Test population and manual job coding

To evaluate the three automated coding systems, we used the lifetime occupational histories reported by subjects from the AsiaLymph study (Friesen *et al.*, 2016). AsiaLymph is a large-scale hospital-based case-control study of lymphoma and leukaemia in Eastern Asia, including 13 253 subjects with over 36 000 reported job records coded in ISCO-88.

Details of the data collection and coding process have been described before (Ge, 2021). Briefly, local research teams interviewed all eligible subjects in person within 48 h of recruitment using a computer-assisted personal interview (CAPI). The interview section on occupational history collected information on all jobs ever held for one year or more by participants and consisted of 15 core questions that included employer name, type of products/services provided by the employer, job title, job tasks, three exposure screening questions, plus numeric information for the starting and ending age/year for each job. All occupational history questions in English from the CAPI system are available in Supplementary Figs. 1–4. For each job, participants were also asked structured, exposure-oriented questions from one of 23 study-specific modules that were selected by an automated algorithm that used a combination of keyword search of the occupational history and answers to three exposure screening questions, as described elsewhere (Friesen *et al.*, 2016).

Job coders used the job description information (in Chinese), but not the screening or module questions, to code jobs into the 5th Revision of Standard Occupational Classification System of the Republic of China (SOCSROC)—the Chinese coding system, which has identical coding structures with ISCO-88. All job codes were assigned independently by two study centres, and discordant assignments and the crosswalk to ISCO-88 were individually reviewed and resolved by a third coder (CG). Most codes (72%) were assigned to the most detailed 4-digit coding level. Inter-coder agreements ranged from 51 to 77% depending on job code digit level and study centres (Ge, 2021). The free-text job description information was professionally translated from traditional and simplified Chinese into English.

Automatic codes and crosswalk

The automatic codes based on the translated job descriptions were produced following the instructions from each coding tool. As listed in Table 1, automatically assigned job code outputs were in ISCO-08 and NOC 2016, while the manually assigned job codes were only available in ISCO-88. ISCO-88 and ISCO-08 have similar coding structures (4-digits code representing major, sub-major, minor, and unit groups), but ISCO-08 has more detailed job codes in sub/minor/unit groups (e.g. ISCO-08 has 436 codes in unit groups versus 390 for ISCO-88) (ILO, 2012). NOC 2016 shares a similar conceptual framework with ISCO-08 and has the same 4-digit hierarchical structure (10 broad occupational categories, 40 major groups, 140 minor groups, and 500 unit groups) (https://www.statcan.gc.ca/en/subjects/standard/noc/2016/introduction).

To ensure a reliable comparison of performance, we considered the original manual codes in ISCO-88 as the reference and crosswalked all code outputs to ISCO-88. AUTONOC output codes were crosswalked from NOC 2016 to ISCO-08 based on the official concordance table of NOC 2011—ISCO-08 (available at www.statcan.gc.ca/en/subjects/standard/noc/2011/noc2011-isco2008; NOC 2016 and NOC 2011 share the same coding structure). The crosswalked ISCO-08 codes from AUTONOC were then, together with the outputs from CASCOT and LabourR, crosswalked from ISCO-08 to ISCO-88 using the official ILO crosswalk instructions (www.ilo.org/public/english/bureau/stat/isco/docs/correspondence08.docx). All one-to-many matches (e.g. one ISCO-08 code corresponds to multiple ISCO-88 codes) were resolved by selecting, among the many matches, the job codes that were most common in the AsiaLymph test data. We provided the percentages of one-to-many matches for all crosswalk processes in Supplementary Table S1.

In addition, to assess the reliability of the crosswalk process, we also crosswalked manual codes from ISCO-88 to ISCO-08 based on the official crosswalk instruction and manual review for the one-to-many matches (53.4% of job records). We then calculated the per cent agreement based on the ISCO-08 codes and compared the agreement against the results from ISCO-88.

Performance assessment

Job coding agreement

Per cent agreement was compared between each automatically assigned job code against the manually assigned one as the reference. The comparisons were made with job codes in each of the four-digit levels (i.e. 1-digit major groups; 2-digit sub-major groups; 3-digit minor groups; 4-digit unit groups) in ISCO-88.

To examine how the coding tools performed across different occupation categories, we also compared coding agreement across the 20 most common 2- and 3-digit level occupation codes in the reference data. For CASCOT outputs, we additionally examined the 4-digit level agreement with reference codes at different confidence score cut-off points.

Reliability of exposure assignment

To study the potential impact of job coding disagreement in exposure assessment of different occupational exposures, we linked all the ISCO-88 jobs codes with ALOHA+ JEM (Skorge *et al.*, 2009), which includes assessment for 12 exposures (i.e. biological dust, mineral dust, gas fumes, vapours/gases/dust/fumes, all pesticides, herbicides, insecticides, fungicides, aromatic solvents, chlorinated solvents, other solvents, metals). We also assessed the exposure assignment of electric injury by linking the SHOCK-JEM (Huss *et al.*, 2013) at the 3-digit level of ISCO-88. Both JEMs use exposure scales of 0 (no exposure), 1 (low exposure), and 2 (high exposure). The exposure levels were set at 0 for entries with missing coding outputs from the automatic coding tools.

For each exposure, we dichotomised exposure status into two categories (unexposed = 0 versus exposed = 1 and 2) and calculated the prevalence-adjusted and bias-adjusted kappa (PABAK), which indicates the agreement between the exposure assignments from the automatic codes and that from the reference codes. PABAK overcomes the limitations of conventional Cohen's kappa statistics by accounting for the prevalence effect, bias effect, and unbalanced marginal total effect (Byrt *et al.*, 1993). We used the standard kappa descriptive scale (<0.00 = poor agreement, 0.00–0.20

= slight agreement, 0.21–0.40 = fair agreement, 0.41–0.60 = moderate agreement, 0.61–0.80 = substantial agreement, and 0.81–1.00 = almost perfect agreement) to interpret the agreement from this study (Landis and Koch, 1977). We provide the prevalence of exposed jobs according to each tool and the manual coding in Supplementary Table S2.

All analyses were conducted in R (version 4.0.5) (Team, 2020). PABAK scores were calculated with the *epiR* package (Stevenson *et al*., 2020).

Results

A total of 36 175 occupations reported by AsiaLymph study subjects were included in the study. A proportion of 0.3% of jobs could not be manually coded. The proportions of uncoded occupations for CASCOT, LabourR, and AUTONOC were 3.0%, 3.3%, and 0.5%, respectively. We presented the distribution of the missing coding by each occupation sector (1-digit code) in the supplementary Table S3.

Job coding assessment

The per cent agreement for automatically versus manually assigned codes in ISCO-88 ranged from 17.7 to 26.0% at the 4-digit level and from 43.8 to 58.1% at the 1-digit level (Table 2). Agreements generally increased when comparisons move towards higher, more aggregate levels of job coding (i.e. 1/2-digit). Among the three automatic coding tools, CASCOT produced job codes that had the highest agreement with the manual codes on all coding granularity levels, followed by AUTONOC (Table 2). The coding agreement compared in ISCO-08 showed little difference (<5%) to the agreement in ISCO-88 (Supplementary Table S4).

Stratified per cent agreement by job sector shows varying degrees of agreement, where CASCOT shows higher 4-digit coding agreement across most sectors than the other coding tools (Table 3). Among the nine job sectors, “Service workers and shop and market sales workers” and “Professionals” indicate much better agreement than the agreement in other sectors (agreement ranges: 37.7–54.5% and 30.4–46.0%, respectively). All three tools performed poorly for “Legislation, senior officials and managers” sector (agreement range: 4.5–11.6%).

Table 4 presents more detailed coding agreement of top three ISCO-88 4-digit codes from each job sector. All the three tools coded “office clerk” (ISCO-88: 4100) and “Market gardeners and crop growers” (6110) differently. AUTONOC and CASCOT tended to code those to 4190—“Other office clerks” (44.5% and 67.4% proportionally), and LabourR mainly coded the office clerk codes (4100) as “Business services agents and trade brokers” (3420). For “Market gardeners and crop growers” (6110), CASCOT and LabourR coded 68.9% and 60.2% of those “6110” codes to “6130” (Market-oriented crop and animal producers) instead.

Agreement of most codes varies substantially among the three coding tools. For example, AUTONOC had markedly better agreement for a similar code “Field crop and vegetable growers” (6111) (55.7%), but coded all “Market-oriented crop and animal producers” (6130) (a similar job code to 6111) differently from the manual codes. CASCOT and LabourR, on the other hand, show better agreement for “Market-oriented crop and animal producers” (6130) (81.4% and 84.1%, respectively), but very low agreement for “Field crop and vegetable growers” (6111). We presented the per cent agreement of 20 most common

Table 2. Per cent agreement on four ISCO-88 coding levels between automatically assigned codes from different coding tools versus manually assigned codes.

ISCO-88*	AUTONOC (%)	CASCOT (%)	LabourR (%)
1-digit	48.8	58.1	43.8
2-digit	42.2	51.4	37.2
3-digit	30.8	33.3	22.9
4-digit	20.3	26.0	17.7

*1-digit: major groups; 2-digit: sub-major groups; 3-digit: minor groups; 4-digit: unit groups.

Table 3. Percent agreement (%) between automatically assigned codes from three coding tools versus manually assigned codes on 4-digit ISCO-88 level, by job sector.

Job sector (by major group)	AUTONOC (%)	CASCOT (%)	LabourR (%)
Legislators, senior officials, and managers	6.6	11.6	4.5
Professionals	42.9	46.0	30.4
Technicians and associate professionals	12.7	16.0	12.9
Clerks	11.2	17.8	15.5
Service workers and shop and market sales workers	46.4	54.5	37.7
Skilled agricultural and fishery workers	10.8	9.6	6.0
Craft and related trades workers	17.3	27.8	15.5
Plant and machine operators and assemblers	15.8	27.9	19.9
Elementary occupations	20.7	22.7	13.2

Table 4. Percent agreement (%) on 4-digit level coding for the top 3 codes by each occupation sector.

Job sector (by major group)	4-digit code[*]	Job title	AUTONOC	CASCOT	LabourR
Legislators, Senior Officials, And Managers	1210	Directors and chief executives (*n* = 930)	1.5	2.5	1.6
	1222	Production and operations department managers in manufacturing (*n* = 89)	4.5	20.2	9.0
	1224	Production and operations department managers in whole sale and retail trade (*n* = 115)	10.4	56.5	3.5
Professionals	2221	Medical doctors (*n* = 273)	65.2	83.5	0.0
	2320	Secondary education teaching professionals (*n* = 305)	2.6	3.3	82.0
	2331	Primary education teaching professionals (*n* = 384)	79.2	68.8	1.6
Technicians And Associate Professionals	3415	Technical and commercial sales representatives (*n* = 527)	1.9	6.6	26.0
	3433	Bookkeepers (*n* = 613)	8.2	4.4	5.9
	3439	Administrative associate professionals not elsewhere classified (*n* = 1762)	9.1	0.5	2.4
Clerks	4100	Office clerks (*n* = 806)	0.0	0.0	0.0
	4111	Stenographers and typists (*n* = 561)	0.9	6.4	4.1
	4131	Stock clerks (*n* = 514)	5.4	6.8	50.6
Service Workers And Shop And Market Sales Workers	5122	Cooks (*n* = 764)	54.8	66.0	57.7
	5123	Waiters, waitresses, and bartenders (*n* = 633)	39.2	38.9	40.4
	5220	Shop salespersons and demonstrators (*n* = 1887)	51.4	50.8	35.8
Skilled Agricultural And Fishery Workers	6110	Market gardeners and crop growers (*n* = 2551)	0.0	0.0	0.0
	6111	Field crop and vegetable growers (*n* = 784)	55.7	5.4	0.0
	6130	Market-oriented crop and animal producers (*n* = 295)	0.0	81.4	84.1
Craft And Related Trades Workers	7212	Welders and flamecutters (*n* = 259)	49.8	69.9	55.2
	7233	Agricultural- or industrial-machinery mechanics and fitters (*n* = 391)	24.0	32.0	18.4
	7241	Electrical mechanics and fitters (*n* = 241)	19.9	19.5	0.0
Plant And Machine Operators And Assemblers	8211	Machine-tool operators (*n* = 536)	31.0	33.4	29.3
	8263	Sewing machine operators (*n* = 479)	27.6	29.2	0.0
	8322	Car, taxi, and van drivers (*n* = 607)	22.1	81.7	75.9
Elementary Occupations	9132	Helpers and cleaners in offices, hotels, and other establishments (*n* = 480)	50.0	7.7	12.3
	9312	Construction and maintenance labourers: roads, dams, and similar constructions (*n* = 399)	23.8	12.8	18.0
	9322	Hand packers and other manufacturing labourers (*n* = 596)	28.0	33.1	12.4

[*] Three most common 4-digit ISCO-88 codes from each job sector, based on the distribution of manual codes.

3-digit codes by the three tools in the Supplementary Table S5.

For CASCOT outputs, agreement with the reference (manual) codes increased with higher confidence scores (Table 5). The output codes with scores higher than 40 (66.3% of all output) showed slightly better agreement than the overall agreement (28.7% versus 26.0%). The per cent agreement reached 43.2% when the codes had confidence scores over 80, yet the proportion of such codes was small (22.2%). For the test dataset, the median of the CASCOT confidence scores was 54 [interquartile range (IQR) 39–78].

Reliability of exposure assignment

PABAK coefficients for the JEM-based exposure levels were in the range of 0.42–0.94, and the median values ranged from 0.69 to 0.78 (Table 6). CASCOT had the highest overall exposure agreement across 12 exposures, with a medium PABAK of 0.78 (IQR: 0.66–0.89, indicating substantial to almost perfect agreement) for ALOHA+ JEM exposures. Both CASCOT and LabourR had the highest PABAK coefficient of 0.94 for "Herbicides" exposure. For CASCOT and LabourR, exposure assignments to "All pesticides", "Herbicides", "Insecticides", and "Fungicides" had an almost perfect agreement with the assignment based on manual job codes (PABAK > 0.80). For exposure assessment on electric shock, CASCOT remained the best performer (PABAK: 0.67, substantial agreement), followed by AUTONOC (0.56, moderate agreement).

Table 5. Agreement of codes stratified with different confidence score cut-off points in ISCO-88 system for CASCOT.

	Confidence score	*n* (%)	Number of matched codes with reference	Agreement (%)
CASCOT	<40	12 190 (33.7%)	2529	20.7
	≥40	23 985 (66.3%)	6874	28.7
	≥60	15 583 (43.1%)	5438	34.9
	≥80	8017 (22.2%)	3465	43.2
	0–100 (all codes)	36 175	9424	26.0

Discussion

We provided an overview of three existing automatic job coding tools and assessed the performance of job coding and JEM-assigned exposure agreement for the tools with job histories translated into English from the AsiaLymph study. Agreement on job coding was generally low for the three tools being compared (less than 26.0% for exact match of 4-digit ISCO-88 codes), but a higher degree of exposure agreement was found. Based on our testing dataset AsiaLymph, CASCOT had a higher overall agreement with human reviewers for all four level of the ISCO-88 coding system and had a higher overall agreement for all the assessed exposures in this study.

Poor job coding agreement can partly be explained by limitations in algorithm performance, particularly in separating similar words with different contextual meanings in occupation classification. In our analysis, although the job codes in the minor group "Market gardeners and crop growers" (6110) were coded into "Market-oriented crop and animal producers" by CASCOT and LabourR, this title still shares some keywords such as "market" and "crop". While poor agreement in job coding was found for the three tools on the 3/4-digit level, a higher level of agreement overall was found for most sub-major groups when the comparison was made on the 2-digit level (in Supplementary Table

Table 6. Prevalence-adjusted and bias-adjusted Kappa (PABAK) coefficients for various exposures from ALOHA+ and SHOCK-JEM-based on automatically versus manually assigned ISCO-88 job codes.

JEM	Exposure	AUTONOC	CASCOT	LabourR
ALOHA +	Biological dust	0.46	0.69	0.60
	Mineral dust	0.51	0.64	0.48
	Gas fumes	0.53	0.59	0.42
	Vapours, gases, dust, and fumes	0.59	0.64	0.43
	All pesticides	0.72	0.86	0.81
	Herbicides	0.79	0.94	0.94
	Insecticides	0.77	0.90	0.92
	Fungicides	0.74	0.90	0.83
	Aromatic solvents	0.64	0.76	0.70
	Chlorinated solvents	0.79	0.80	0.77
	Other types of solvents	0.66	0.73	0.68
	Metals	0.75	0.79	0.70
	Median (IQR)	0.69 (0.54–0.77)	0.78 (0.66–0.89)	0.70 (0.51–0.83)
SHOCK[*]	Electric shock	0.56	0.67	0.49

[*] Linkage of codes was conducted on the 3-digit level for SHOCK-JEM.
IQR, Interquartile range. The 95% confidence intervals were shown in Supplementary Table S7.

S6), suggesting their capacity to sufficiently code some sub-major occupations.

Another possible explanation for the poor performance is the limited input information. Most of the tools listed in Table 1 rely only on job titles, and only a few include other information such as job tasks and industry. The limited input information these coding tools rely on is in contrast with a much wider assortment of job information for trained human job coders to perform the task. Such an effect may be reflected by the much lower coding agreement from the three tools (17.7–26.0%, Table 2) when compared with the inter-rater agreement for this dataset [~51% (Ge, 2021)] and for other testing datasets [e.g. 60% (Kennedy *et al.*, 2000) and 36–68% (Kromhout and Vermeulen, 2001; Koeman *et al.*, 2013)]. For our test dataset, manual job coders used job title, job task, industry, service/product provided, calendar year, as well as full occupational histories for all subjects to assign job codes. Job coders were also trained to perform online searches for company names when there were ambiguous names. The additional information would provide extra contextual meanings to the free-text job description, especially when job title descriptions are ambiguous (e.g. factory worker; labourer). It would be possible that the coding performance from the job coding algorithms could be significantly improved (or even better than manual coding) provided that the tool could incorporate a large amount of data inputs to fill the knowledge gaps while mimicking the way humans process those data with some state-of-the-art algorithms. Although, for such purpose, it can be difficult to obtain large high-quality datasets with all relevant variables for training and testing.

While the three job coding tools were selected based on similarities in job classification, we still explored whether the information loss from the crosswalk process could have introduced coding errors by comparing the coding agreement in ISCO-88 and ISCO-08 (Supplementary Table S4). Results for the two sets of comparison were very similar (difference in per cent agreement <5%), suggesting that the applied automatic crosswalk procedure was unlikely to introduce a significant amount of error to our main findings.

For AUTONOC and CASCOT, marked differences were found between the previously reported coding agreement and those based on this study. For CASCOT coding outputs with confidence scores over 40, the agreement difference between the reported 80% and 28.7% from this study could be explained by different job distribution in the training dataset (Warwick Institute for Employment, 2018). Similarly, the differences from AUTONOC might be due to the difference in occupation code frequency in our testing data as compared with the training dataset (Garcia *et al.*, 2021).

The agreement was higher when comparisons were made at the job-exposure level than the agreement at job coding level. A better overall agreement (indicated by much higher PABAK) was found for the 13 exposures. Similar results have also been reported by other studies (Kennedy *et al.*, 2000; Kromhout and Vermeulen, 2001; Koeman *et al.*, 2013; Russ *et al.*, 2016), because many job codes share the same exposure category across different occupational groups in JEMs. Furthermore, job classification systems used in epidemiology studies were not originally designed for occupational exposure assessments. It can be argued that disagreement in job coding by automatic job coding tools does not necessarily lead to disagreement on exposure assignments from JEMs; tools that are intendedly developed for epidemiological applications could therefore consider the metrics of exposure assignment as one important criterion of evaluation.

One distinct advantage of the current analysis is the use of a large testing dataset with high-quality manually assigned ISCO-88 job codes in AsiaLymph. Dedicated efforts were made to produce the manual job codes by two centres in Asia (Ge, 2021). The large number of job codes (>36 000) provide a realistic comparison of job coding performance in the context of large-scale occupation health studies for the included coding tools. Additionally, it was found that higher scores of the outputs lead to better agreement with the manual codes, however, the proportion of such "more confident" outputs (i.e. job codes with higher confidence scores) is small. The choices of the cut-off points, however, are largely study-specific and are highly dependent on the trade-off between the level of coding agreement required and the cost of workload for the rest of the "not confident" codes to be expert-reviewed.

Our work also has limitations. Firstly, the results of comparison from the current study may not be generalisable to other cohort data as job histories from East Asia may not be representative of job distributions in other regions. For instance, job histories from Western Europe may contain a smaller proportion of jobs in agriculture, which would likely alter the comparison results among the three automatic coding tools. Secondly, there may be a difference in how people understand and describe their job across different regions and cultural backgrounds. On close review of the AsiaLymph jobs, study investigators found that, similar to other studies, the additional information reported in the tasks, employer name, and products made/services provided crucial information necessary for accurate coding; however, this information was not fully used by the coding systems. Thirdly, the professional translations of the free-text job descriptions were undertaken to aid exposure assessors, alongside detailed exposure-oriented questions, but may not reflect the terminology more commonly observed in occupational data collected in English-speaking populations. Additionally, although the similar results from coding agreement in ISCO-88 and ISCO-08 indicate the validity of crosswalk processes, the performance from none of the three tools was compared in their original coding output systems. This is because no manual codes in ISCO-08 were available for the testing dataset. We acknowledge possible errors introduced by the crosswalk process may not be sufficiently assessed by comparing the coding agreement difference alone. Finally, the number of tools being compared was limited, largely because of the various coding systems of their outputs and the absence of reliable crosswalks between these systems (e.g. ISCO and US SOC) to make the comparison possible.

Moreover, as mentioned in Russ *et al.* (2016), our evaluation of the coding "performance" does not necessarily reflect the validity of job code assignment. Despite huge efforts, the inter-coder agreement of the manual codes in Asialymph ranged from 51% (on 4-digit level) to 77% (on 1-digit level) (Ge, 2021). It is important to reiterate that job coding is inherently a subjective and variable activity, insofar as true gold standards for job codes rarely exist outside of specific examples in classification documents. The comparison of performance in this analysis reflects the capacity of how the coding tools can reproduce the result from human coders, while the true validity of codes remains difficult to assess.

Nevertheless, automatic coding tools have clear advantages in cost and time efficiency over human counterparts, even if fully automated coding performance is still relatively poor compared to human experts. The work to manually code the jobs in the test dataset lasted approximately four years across multiple study centres. Significant time and cost would be saved if an automatic coding tool could pre-screen and reliably code even a relatively small proportion of all jobs. For the remaining part, the automatic coding tool could provide a set of options to choose from, which is already provided by CASCOT and AUTONOC to make the manual job coding process more efficient. Another advantage of automated coding is high intra-tool reliability, whereas human job coders might assign somewhat different job codes to the same set of jobs on different occasions. Therefore, any systematic misclassification by automated systems can be identified and corrected en masse, by recoding with a new version of a coding algorithm. Moreover, systematic recoding is trivial in terms of computer time.

Some algorithms can also output justifications of choices or lists of multiple candidates which can be verified by manual coders. For example, AUTONOC has also recently introduced a portal for study managers to email study participants each a unique URL to self-code their occupation and select from candidates' occupations provided by the algorithm (Bao *et al.*, 2020; Garcia *et al.*, 2021).

It is also worth noting that a large amount of "legacy" occupational information remains underutilized in various cohorts, registries, and records due to the lack of resources for job code assignments. It is fair to say that even though automatic coding tools might not fully replace human coders, these tools still have significant potential to be widely adopted for prescreening in large sets of general population data.

Conclusions

We presented an overview of several automatic job coding tools and compared the performance of three tools (AUTONOC, CASCOT, and LabourR) in job coding and JEM-assessed exposures based on translated lifetime job histories collected from the AsiaLymph study. Overall, the exact agreement on job coding was low for all three tools. In this comparison, CASCOT outperformed the other two in terms of the better agreement in job coding and exposure assignment. This work not only indicates the importance of study-specific evaluations prior to the use of coding tools but also signifies the need for improvements in evaluated automatic coding tools.

Acknowledgements

We thank Prof Peter Elias (University of Warwick) for his useful comments on the manuscript.

Funding

This work was conducted within the scope of the Exposome Project for Health and Occupational Research (EPHOR), funded by the European Union's Horizon 2020 research and innovation programme under grant agreement number 874703. The AsiaLymph study was funded by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, NCI, NIH.

Conflict of interest

None of the authors has any conflicts of interest in this work.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

SUPPLEMENTARY DATA

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

References

Bao H, Baker CJO, Adisesh A. (2020) Occupation coding of job titles: iterative development of an Automated Coding Algorithm for the Canadian National Occupation Classification (ACA-NOC). *JMIR Form Res*; **4**: e16422.

Burstyn I, Gustafson P, Pintos J *et al*. (2018) Correction of odds ratios in case-control studies for exposure misclassification with partial knowledge of the degree of agreement among experts who assessed exposures. *Occup Environ Med*; **75**: 155–9.

Burstyn I, Slutsky A, Lee DG *et al*. (2014) Beyond crosswalks: reliability of exposure assessment following automated coding of free-text job descriptions for occupational epidemiology. *Ann Occup Hyg*; **58**: 482–92.

Byrt T, Bishop J, Carlin JB. (1993) Bias, prevalence and kappa. *J Clin Epidemiol*; **46**: 423–9.

Friesen MC, Lan Q, Ge C *et al*. (2016) Evaluation of automatically assigned job-specific interview modules. *Ann Occup Hyg*; **60**: 885–99.

Ge C. (2021) *Occupational exposure assessment in the general population: improvements, innovations, and impact*. The Netherlands: Utrecht University. ISBN: 9789083129556.

Huss A, Vermeulen R, Bowman JD *et al*. (2013) Electric shocks at work in Europe: development of a job exposure matrix. *Occup Environ Med*; **70**: 261–7.

ILO. (2012) International Standard Classification of Occupations 2008 (ISCO-08): Structure, group definitions and correspondence tables: International Labour Office. ISBN 978 92 2 125952 7

Kennedy SM, Le Moual N, Choudat D *et al*. (2000) Development of an asthma specific job exposure matrix and its application in the epidemiological study of genetics and environment in asthma (EGEA). *Occup Environ Med*; **57**: 635–41.

Koeman T, Offermans NS, Christopher-de Vries Y *et al*. (2013) JEMs and incompatible occupational coding systems: effect of manual and automatic recoding of job codes on exposure assignment. *Ann Occup Hyg*; **57**: 107–14.

Kogevinas M, Schlünssen V, Mehlum IS *et al*. (2020) The OMEGA-NET International inventory of occupational cohorts. *Ann Work Expo Health*; **64**: 565–8.

Kouretsis A, Bampouris A, Morfiris P, Papageorgiou K. (2020) labourR: classify multilingual labour market free-text to standardized hierarchical occupations. Available from: https://CRAN.R-project.org/package=labourR. Accessed 23 July 2021.

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

Landis JR, Koch GG. (1977) The measurement of observer agreement for categorical data. *Biometrics*; **33**: 159–74.

Patel MD, Rose KM, Owens CR *et al*. (2012) Performance of automated and manual coding systems for occupational data: a case study of historical records. *Am J Ind Med*; **55**: 228–31.

Russ DE, Ho KY, Colt JS *et al*. (2016) Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies. *Occup Environ Med*; **73**: 417–24.

Savic N, Bovio N, Gilbert F *et al*. (2022) Procode: a machine-learning tool to support (Re-)coding of free-texts of occupations and industries. *Ann Work Expo Health*; **66**: 113–8.

Skorge TD, Eagan TM, Eide GE *et al*. (2009) Occupational exposure and incidence of respiratory disorders in a general population. *Scand J Work Environ Health*; **35**: 454–61.

Suarez Garcia CA, Adisesh A, Baker CJ. (2021) S-464 automated occupational encoding to the Canadian National Occupation classification using an ensemble classifier from TF-IDF and Doc2Vec Embeddings. *Occup Environ Med*; **78**: A161.

Stevenson M, Sergeant E, Nunes T *et al*. (2020) epiR: tools for the analysis of epidemiological data. Available from: https://CRAN.R-project.org/package=epiR. Accessed 11 June 2021.

R Core Team. (2020) R: a language and environment for statistical computing. Available from: https://www.R-project.org. Accessed 21 June 2022.

Warwick Institute for Employment R. (2018) Cascot: Computer Assisted Structured Coding Tool. [serial online] 2018. Available from: https://warwick.ac.uk/fac/soc/ier/software/cascot/details/. Accessed 7 July 2020.