Accepted: 13 November 2018
DOI: 10.1002/ajim.22928

RESEARCH ARTICLE

WILEY AMERICAN JOURNAL OF INDUSTRIAL MEDICINE

Efficiency of autocoding programs for converting job descriptors into standard occupational classification (SOC) codes

Skye Buckner-Petty MPH | Ann Marie Dale PhD, OTR/L | Bradley A. Evanoff MD, MPH

Division of General Medical Sciences, Washington University School of Medicine, St. Louis, Missouri

Correspondence
Skye Buckner-Petty, MPH, Division of General Medical Sciences, Washington University School of Medicine, 4523 Clayton Avenue, Campus Box 8005, St. Louis, MO 63110.
Email: skye.skye@wustl.edu

Funding information
National Institute for Occupational Safety and Health, Grant number: R01OH011076

Background: Existing datasets often lack job exposure data. Standard Occupational Classification (SOC) codes can link work exposure data to health outcomes via a Job Exposure Matrix, but manually assigning SOC codes is laborious. We explored the utility of two SOC autocoding programs.

Methods: We entered industry and occupation descriptions from two existing cohorts into two publicly available SOC autocoding programs. SOC codes were also assigned manually by experienced coders. These SOC codes were then linked to exposures from the Occupational Information Network (O*NET).

Results: Agreement between the SOC codes produced by autocoding programs and those produced manually was modest at the 6-digit level, and strong at the 2-digit level. Importantly, O*NET exposure values based on SOC code assignment showed strong agreement between manual and autocoded methods.

Conclusion: Both available autocoding programs can be useful tools for assigning SOC codes, allowing linkage of occupational exposures to data containing free-text occupation descriptors.

KEYWORDS
industry and occupation coding, job exposure matrix, NIOCCS, O*NET, SOCcer

1 | INTRODUCTION

Analyzing occupational risk factors is of great interest in epidemiological research. Working Americans spend a large portion of their day at the workplace,[1] therefore it is important to focus on determinants of health that are related to work and workplace behaviors. However, classifying jobs and identifying occupational risk factors can be challenging, especially when using pre-existing data. When this occupational information is available, it is most commonly in the form of free-text responses to open-ended questions regarding industry, job titles, and/or job tasks. In order to analyze these data, it is necessary to assign Standard Occupational Classification (SOC) codes to the free-text entries. The SOC system is used by federal statistical agencies in the United States for the purpose of collecting, calculating, and disseminating data. These SOC codes have a hierarchical structure. The first two digits indicate a classification into one of 23 "major groups." There are 97 "minor groups" represented by the first four digits, and 461 "broad occupations" represented by the first five digits. Finally, six digits represent a "detailed occupation" classification, of which there are 840.[2] For research purposes, any of these classification levels can be used for studying associations between occupation and health outcomes and behaviors. The greater the number of digits indicated in the SOC codes, the higher the level of homogeneity of the characteristics in jobs represented by the code.

Institution at which the work was performed: Washington University School of Medicine, St. Louis, MO.

SOC codes can also be used to assign various job-level exposure estimates by linking job titles to job exposure matrices (JEMs). JEMs provide a source to obtain job related exposure data for existing studies that lack such data. JEMs are commonly used for studying chemical, electromagnetic, and noise exposures. However, JEMs can be useful to the study of a wide range of health outcomes. There has been recent interest in applying JEMs to study the effects of workplace physical demands on musculoskeletal disorders, pregnancy outcomes, hernias, and cardiovascular disease.[3–7] Such job-related data can be obtained through The Occupational Information Network (O*NET)[8] provided by the U.S. Department of Labor. O*NET is a publicly available database that contains a variety of descriptors of physical and mental demands related to SOC codes. Several recent studies have used O*NET-based physical exposure estimates to evaluate relationships between workplace exposures and chronic disease outcomes.[9–11] One recent study showed the reliability of some O*NET-based force and repetition variables of the hand and wrist compared to similar directly observed measures, and demonstrated their ability to predict incident carpal tunnel syndrome.[3]

Manually assigning SOC codes is a time-consuming process; in large studies, this method can be infeasible. In addition, manual coding is not conducive to research reproducibility. Two publicly available autocoding programs have been created to alleviate this burden from researchers and facilitate large-scale studies that require occupation data (the US National Institute for Occupational Safety and Health has developed the NIOSH Industry and Occupation Computerized Coding System (NIOCCS) Version 3, and The National Institutes of Health provided Standardized Occupation Coding for Computer-assisted Epidemiological Research (SOCcer) Version 1). NIOCCS has been previously reviewed by Schmitz et al[12] and Weiss et al[13] to show feasibility of use, however, additional testing in various datasets is necessary to demonstrate performance capabilities. This program has recently been used to categorize occupational information in Behavioral Risk Factor Surveillance System (BRFSS) data.[14,15] Details of SOCcer are described by Russ[16] along with preliminary testing. However, to our knowledge, SOCcer has not been formally tested by researchers unaffiliated with the development team. Furthermore, no studies have compared the two programs using data from published cohorts. The purpose of this study was to test and compare the performance of both job title autocoding programs using data from two existing epidemiological studies. We compared agreement between SOC codes from the coding programs to that of expert manual coders. We also compared agreement between autocoding and manual coding on the subsequent exposure values obtained when using assigned SOC codes to extract six relevant occupational physical exposures from O*NET.

2 | METHODS

2.1 | Study population

Two cohorts from previously completed studies were used for the current study. The secondary data from these cohorts was de-identified; IRB review was not required. Cohort 1 consisted of pooled data from six prospective studies of workplace risk factors for upper extremity musculoskeletal disorders. Details of the pooled cohort have been described elsewhere.[17,18] Briefly, study participants were fulltime male and female employees, 18 years of age or older, who worked in a variety of industries including manufacturing, production, service, and construction. In total, 4321 workers were recruited across the six study sites and followed between 2001 and 2010.

Cohort 2 consists of data from a cross-sectional, telephone-based survey that was conducted as part of the Supports at Home and Work for Maintaining Energy Balance (SHOW-ME) study. The SHOW-ME study was designed to examine the associations between residential and worksite environmental and policy influences and energy balance behaviors and outcomes. There were 2015 participants in this cohort who were 21-65 years of age, and employed at least 20 h per week. Further details of the SHOW-ME study are described elsewhere.[19]

2.2 | Manual job coding

From cohort 1, information about each worker's current job was collected at baseline and follow-up including job title, company name, and work-related tasks. The last known job for each subject was used for this study as this provided a greater number of job titles than baseline data. In cohort 2, participants were asked to provide the name and street address of their primary workplace. They then were asked what kind of work they do (occupation) and in what type of business or industry. Using the available information from both cohorts, SOC codes were assigned to each worker using the job title selection feature provided by the O*NET online database[8] and selecting the occupational code that best matched the primary tasks and employer information. Standard Industrial Classification (SIC) codes and National American Industry Classification System (NAICS) codes were assigned in a similar fashion after using ReferenceUSA[20] to look up the employers of each worker. SOC, SIC, and NAICS codes were independently assigned by two raters, with differences resolved by consensus. O*NET assigns job exposure information for SOC codes at the detailed occupation level. Since linking free text job titles to O*NET physical exposures was the primary motivation for assigning SOC codes in both cohorts, the availability of exposure values influenced the coding decisions. Only subjects with sufficient job information that allowed for successful manual SOC, SIC, and NAICS code assignments were included in the current study.

2.3 | NIOCCS autocoding

NIOCCS is capable of assigning SOC codes using a combination of industry (input as free-text or NAICS codes) and occupation free-text. The SOC codes presented in the output can range in level of detail from 2 to 6 digits. Each code is provided with a score to indicate the level of confidence of the assignment. These scores will range from 90% to 100%; codes with scores less than 90% are not included in the output. Self-reported occupation free-texts and manually assigned NAICS

codes from both cohorts were input into the NIOCCS system. This was then repeated except omitting the NAICS codes.

2.4 | SOCcer autocoding

SOCcer codes job descriptions to the SOC 2010 classification system as described in Russ. It allows for batch input of occupation, SIC code, and job task; partial information is allowed. For each row of job-descriptors, the system outputs 10 six-digit SOC codes, each of which is assigned a confidence score representing the estimated probabilities (computed from logistic regression) that an expert reviewer would have selected that SOC code. For convenience, the paper will refer to scores from SOCcer and NIOCCS with the same terminology, although their values are not directly comparable. Unlike NIOCCS, SOCcer will present SOC code options for every entry regardless of confidence levels. It is the user's responsibility to omit codes with low scores. Self-reported occupation based on free text responses and manually assigned SIC codes from both cohorts were used to produce SOCcer output. This was then repeated except with the SIC codes omitted. The SOCcer produced SOC code with the highest confidence score for each subject was used for analyses. In the presence of ties, the first code presented was used.

2.5 | O*NET exposures

Using the SOC codes assigned to each subject by all three coding methods (manual, SOCcer, and NIOCCS), physical work exposure variables were extracted from O*NET databases. A total of six items describing physical workplace demands were selected from three different O*NET databases (work activities, work context, and work abilities). The selected physical exposure items were those used in past studies of the two cohorts, and included: (i) dynamic strength; (ii) static strength; (iii) handling and moving objects; (iv) time spent making repetitive movements; (v) importance of using computers; and (vi) performing general physical activities. The values for each of these exposures were assigned to the participant's SOC codes. O*NET exposures are classified by eight-digit SOC codes. Since both autocoding programs produce up to six-digit SOC codes, ".00" was attached to each code prior to extracting the O*NET values. SOC codes from NIOCCS that were presented with fewer than 6-digits were not linked to O*NET data.

2.6 | Analyses

Cohen's Kappa statistic was used to measure agreement between SOC codes produced manually and by the autocoding programs. Agreement was computed for six-digit (Detailed Occupation), and two-digit (Major Group) SOC codes. Agreement between NIOCCS and manual codes was assessed with and without inputted NAICS codes. Agreement between SOCcer and manual codes was assessed with and without inputted SIC codes. Furthermore, SOC agreement was assessed at various levels of confidence scores for both programs. We assessed the distributions of the confidence scores to determine appropriate stratifications. We also assessed the overall effectiveness of both programs by computing the percentage of each cohort that was accurately autocoded. Cohen's Kappa was also used to measure agreement directly between NIOCCS and SOCcer with and without inputted industry codes.

Intraclass correlations (ICC) were computed to measure agreement between the O*NET values linked to SOC codes produced manually versus those produced by autocoding programs with industry codes and occupation text included as inputs. The frequency of SOC codes produced by autocoding programs that had available O*NET data was also captured. For interpretation of the kappa and ICC values, we followed the guidelines set by Cicchetti[21] and considered values less than 0.40 as poor; between 0.40 and 0.59 as fair; between 0.60 and 0.74 as good; and between 0.75 and 1.00 as excellent. All analyses were conducted using R version 3.3.2.[22]

3 | RESULTS

The inclusion criteria were met by 1823 subjects (three out of six study sites) in cohort 1 and 1496 subjects in cohort 2. In cohort 1, three of the study sites reported job tasks but not job titles. Additional subjects were dropped from both cohorts due to a missing manual SOC, SIC, or NAICS code assignment. Incomplete manual codes resulted from missing or ambiguous job information. Table 1 shows the frequencies of the 23 SOC major groups (defined by the manual SOC codes). Both cohorts have a wide variety of job types. Cohort 2 includes subjects in all SOC major groups except "Military Specific Occupations." Cohort 1 includes subjects in all SOC major groups except "Military Specific Occupations" and "Farming, Fishing, and Forestry Occupations." Cohort 2 had more job diversity than Cohort 1, but had relatively large proportions in "Office and Administrative Support Occupations" (17.8%), "Education, Training, and Library Occupations" (14.2%), and "Healthcare Practitioners and Technical Occupations" (12.5%). Cohort 1 had a large proportion in "Production Occupations" (33.1%). Cohorts 1 and 2 contained 263 and 322 unique six-digit SOC codes, respectively.

Table 2 shows agreement levels for SOC codes assigned by autocoding programs with manually assigned SOC codes. SOCcer codes were stratified into three categories based on confidence scores: less than 0.2, 0.2-0.4, and greater than 0.4. For both cohorts, roughly a third of the codes fit into each of these strata when SIC codes were included as an input. When SIC codes were omitted, the majority of codes had a confidence score less than 0.2 (cohort 1 = 58.7%; cohort 2 = 62.0%). Confidence scores in NIOCCS ranged from 90 to 100; however more than half of the available scores were 100 for both cohorts, with and without NAICS inputted; so we created two strata: scores equal to 100 versus scores less than 100. NIOCCS did not produce SOC codes for all entries. Most entries were coded when NAICS codes were inputted (cohort 1 = 84.6%; cohort 2 = 79.5%). However, few were coded when NAICS codes were omitted (cohort 1 = 31.4%; cohort 2 = 45.1%).

TABLE 1 Frequency of manually coded SOC major groups from pooled upper extremity study (cohort 1), United States, 2001-2010, and from the show-ME Study (cohort 2), Missouri, 2012-2013

	Cohort 1 (*N* = 1823)		Cohort 2 (*N* = 1496)	
SOC major group	***n***	**%**	***n***	**%**
11-0000 Management Occupations	35	1.92	118	7.89
13-0000 Business and Financial Operations Occupations	39	2.14	73	4.88
15-0000 Computer and Mathematical Occupations	25	1.37	55	3.68
17-0000 Architecture and Engineering Occupations	26	1.43	27	1.80
19-0000 Life, Physical, and Social Science Occupations	21	1.15	27	1.80
21-0000 Community and Social Services Occupations	1	0.05	44	2.94
23-0000 Legal Occupations	3	0.16	18	1.20
25-0000 Education, Training, and Library Occupations	20	1.10	212	14.17
27-0000 Arts, Design, Entertainment, Sports, and Media Occupations	4	0.22	26	1.74
29-0000 Healthcare Practitioners and Technical Occupations	103	5.65	187	12.50
31-0000 Healthcare Support Occupations	41	2.25	39	2.61
33-0000 Protective Service Occupations	16	0.88	23	1.54
35-0000 Food Preparation and Serving Related Occupations	76	4.17	67	4.48
37-0000 Building and Ground Cleaning and Maintenance Occupations	83	4.55	46	3.07
39-0000 Personal Care and Service Occupations	9	0.49	38	2.54
41-0000 Sales and Related Occupations	57	3.13	94	6.28
43-0000 Office and Administrative Support Occupations	169	9.27	266	17.78
45-0000 Farming, Fishing, and Forestry Occupations	105	5.76	0	0.00
47-0000 Construction and Extraction Occupations	156	8.56	9	0.60
49-0000 Installation, Maintenance, and Repair Occupations	40	2.19	32	2.14
51-0000 Production Occupations	604	33.13	47	3.14
53-0000 Transportation and Material Moving Occupations	190	10.42	48	3.21
55-0000 Military Specific Occupations	0	0.0	0	0.00

SOC, standard occupational classification.

Both autocoding programs consistently showed increased agreement with manually assigned SOC codes for higher confidence scores regardless of cohort and inputs. Overall, NIOCCS had excellent agreement with two-digit SOC codes (κ = 0.75-0.82) and fair to good agreement for six-digit SOC codes (κ = 0.41-0.63). SOCcer had fair to good agreement with two-digit SOC codes (κ = 0.55-0.70) and poor to fair agreement with six-digit SOC codes (κ = 0.30-0.41). Within the subset of jobs that were coded by both programs, NIOCCS showed stronger agreement than SOCcer for both cohorts.

NIOCCS and SOCcer had similar levels of effectiveness when industry codes were included as inputs. However, when industry was omitted, NIOCCS's effectiveness drastically decreased. With industry included, NIOCCS's effectiveness ranged from 64.0% to 67.4% for two-digit codes, and 33.6% to 44.4% for six-digit codes. However, with industry omitted, NIOCCS's effectiveness ranged from 24.3% to 37.8% for two-digit codes, and 16.1% to 28.7% for six-digit codes. On the contrary, SOCcer's effectiveness only decreased slightly when industry was omitted. With industry included, SOCcer's effectiveness ranged from 62.4% to 72.3% for two-digit codes, and 31.3% to 41.8% for six-digit codes. With industry omitted, SOCcer's effectiveness ranged from 61.2% to 68.2% for two-digit codes, and 30.6% to 40.5% for six-digit codes.

As previously mentioned, the confidence scores from both programs are not directly comparable. However, Figure 1 shows agreement levels at various confidence score cutoffs such that SOCcer scores and NIOCCS scores are aligned by the proportion of the cohort that is coded. NIOCCS had higher agreement than SOCcer in both cohorts for all comparable confidence levels.

Table 3 shows agreement between the SOC codes produced by both autocoding programs. When industry codes were included as inputs, we observed poor to fair agreement for six-digit SOC codes (κ = 0.37-0.49) and good to excellent agreement for two-digit SOC codes (κ = 0.60-0.75). When industry was omitted, we observed fair to good agreement with six-digit SOC codes and excellent agreement for two-digit SOC codes (κ = 0.77-0.80).

Table 4 shows the agreement of O*NET physical exposure variables for SOC codes from autocoding programs compared to those of manually coded SOC codes. Agreement was good to excellent for most comparisons in both programs. Again, higher confidence levels led to stronger agreement. Consequently, both programs produced

TABLE 2 Agreement between autocoded and manually coded SOC codes from pooled upper extremity study (Cohort 1), United States, 2001-2010, and from the show-ME Study (Cohort 2), Missouri, 2012-2013

				SOC major group (2-digit)				Detailed SOC (6-digit)			
Program	Data source	Input variables	Confidence score	N coded	% coded	Kappa	Effectiveness[a]	N coded	% coded	Kappa	Effectiveness[a]
NIOCCS	Cohort 1	Occupation and NAICS	All (90-100)	1542	84.59	0.76	67.42	1530	83.93	0.41	34.56
			Less than 100	681	37.36	0.70		674	36.97	0.31	
			100	861	47.23	0.80		856	49.96	0.48	
		Occupation	All (90-100)	573	31.43	0.75	24.30	559	30.66	0.52	16.07
			Less than 100	178	9.76	0.61		169	9.27	0.38	
			100	395	21.67	0.81		390	21.39	0.57	
	Cohort 2	Occupation and NAICS	All (90-100)	1189	79.48	0.79	64.04	1141	76.27	0.57	44.39
			Less than 100	403	26.94	0.62		385	25.74	0.40	
			100	786	52.5	0.87		756	50.53	0.66	
		Occupation	All (90-100)	675	45.12	0.82	37.77	669	44.72	0.63	28.74
			Less than 100	214	14.30	0.61		208	13.90	0.45	
			100	461	30.82	0.92		461	30.82	0.71	
SOCcer	Cohort 1	Occupation and SIC	All (0-1)	1823	100.00	0.56	62.42	1823	100.00	0.31	31.32
			Less than 0.2	706	38.73	0.38		706	38.73	0.19	
			0.2 to 0.4	428	23.48	0.59		428	23.48	0.20	
			Greater than .4	689	37.79	0.68		689	37.79	0.49	
			90-100 in NIOCCS	1542	84.59	0.58		1530	83.93	0.33	
		Occupation	All (0-1)	1823	100.00	0.55	61.22	1823	100.00	0.30	30.55
			Less than 0.2	1070	58.69	0.44		1070	58.69	0.18	
			0.2 to 0.4	326	17.88	0.72		326	17.88	0.43	
			Greater than .4	427	23.42	0.64		427	23.42	0.48	
			90-100 in NIOCCS	573	31.43	0.70		559	30.66	0.47	
	Cohort 2	Occupation and SIC	All (0-1)	1496	100.00	0.70	72.33	1496	100.00	0.41	41.84
			Less than 0.2	507	33.89	0.50		507	33.89	0.25	
			0.2 to 0.4	476	31.82	0.77		476	31.82	0.32	
			Greater than .4	513	34.29	0.81		513	34.29	0.65	
			90-100 in NIOCCS	1189	79.48	0.75		1141	76.27	0.48	
		Occupation	All (0-1)	1496	100.00	0.65	68.18	1496	100.00	0.40	40.51
			Less than 0.2	927	61.97	0.56		927	61.97	0.26	

(Continues)

TABLE 2 (Continued)

Program	Data source	Input variables	Confidence score	SOC major group (2-digit)				Detailed SOC (6-digit)			
				N coded	% coded	Kappa	Effectiveness[a]	N coded	% coded	Kappa	Effectiveness[a]
			0.2 to 0.4	249	16.64	0.78		249	16.64	0.58	
			Greater than .4	320	21.39	0.79		320	21.39	0.67	
			90-100 in NIOCCS	675	45.12	0.80		669	44.72	0.52	

NIOCCS, NIOSH industry and occupation computerized coding system; SOC, standard occupational classification; SOCcer, standardized occupation coding for computer-assisted epidemiological research.
[a]Percent of total cohort assigned an autocode that matches the manual code.

some codes that were not linkable to O*NET exposures. NIOCCS was slightly more likely to produce SOC codes that did not have exposure data available in O*NET. For cohort 1, 65.1% of the NIOCCS produces codes versus 75.2% of the SOCcer produced codes were linkable to O*NET exposures. Likewise, for cohort 2, 82.4% of the NIOCCS produced codes versus 86.8% of the SOCcer produced codes were linkable to O*NET exposures.

4 | DISCUSSION

We evaluated the performance abilities of SOCcer and NIOCCS to produce SOC codes from self-reported text entries in comparison to SOC codes manually coded by researchers in two published cohort studies. We also evaluated the agreement of O*NET exposure values linked to the aforementioned SOC codes. Both programs demonstrated to be useful tools. NIOCCS had slightly better agreement with manually coded SOC codes than SOCcer when industry codes were included. However, in the absence of industry codes, SOCcer was much more effective. Overall, both programs performed modestly for six-digit SOC codes, with agreement to manually produced codes ranging from poor to good. However, much stronger agreement was observed for two-digit SOC codes; similarly, strong agreement was observed between O*NET exposures linked to the six-digit SOC codes produced manually versus those produced by the autocoding programs. This shows that even when disagreement was found for six-digit SOC codes, the discrepant coding assignments were quite similar in terms of the job classifications and the job-exposures. Therefore, for most research purposes, these discrepancies are likely to be inconsequential.

The agreement that we observed between SOCcer codes and manually coded SOC codes closely mirrored the results from Russ et al[16] at the 2-digit and the 6-digit levels. A previous evaluation of an earlier version of NIOCCS, by Schmitz and Forst,[12] made comparisons at the two- and four-digit levels. Their findings at the two-digit level were close to our results. The strength of agreement that they found at the four-digit level was weaker than that of the two-digit level, but still stronger than what we found at the six-digit level. This suggests that there is reasonable agreement across various datasets between these two autocoding programs. To our knowledge, no previous papers have evaluated agreement between job-exposures estimates made based on autocoded and manually coded job titles. This latter comparison is important, since many researchers seek to code job titles in order to make estimates of work-related exposures. Our finding of good to excellent agreement between exposure values assigned based on manual coding versus autocoding may encourage more researchers to assign workplace exposures based on job title information in a variety of epidemiology studies that would otherwise lack work exposure data.

Both autocoding programs have additional features that were not evaluated in this study. Deciding which program to use will depend on aligning the data with the available features. NIOCCS accepts industry in the form of free text as well as NAICS codes; while SOCcer requires industry to be included as SIC codes. Often, data will include industries that are not already coded, thus making NIOCCS the appropriate

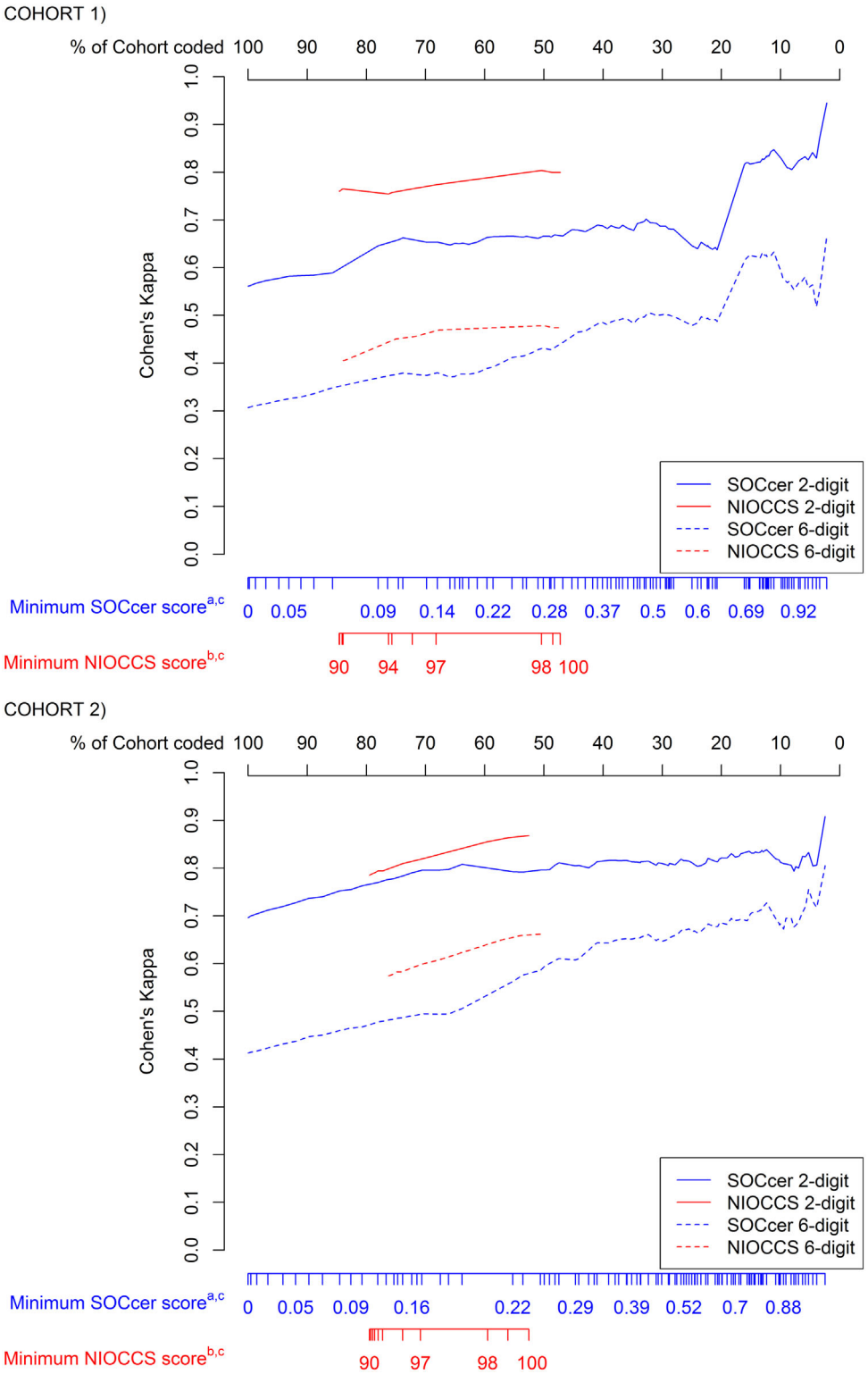

FIGURE 1 Agreement between autocoded and manually coded SOC codes based on minimum allowed confidence scores from pooled upper extremity study (Cohort 1), United States, 2001-2010, and from the show-ME study (Cohort 2), Missouri, 2012-2013. NIOCCS (NIOSH industry and occupation computerized coding system); SOC (standard occupational classification); SOCcer (standardized occupation coding for computer-assisted epidemiological research). a, SOCcer scores range from 0 to 1; b, NIOCCS scores range from 90 to 100 (SOC codes with scores less than 90 are omitted by the program); and c, SOCcer and NIOCCS scores are aligned based on the proportion of the cohort that is autocoded at each minimum threshold. Industry codes and occupation text included for both autocoding programs

TABLE 3 Agreement Between NIOCCS and SOCcer SOC Codes From Pooled Upper Extremity Study (Cohort 1), United States, 2001-2010, and From the Show-ME Study (Cohort 2), Missouri, 2012-2013

		Detailed SOC (6-digit)		SOC major group (2-digit)	
Data Source	Input	N[a]	Cohen's Kappa	N[b]	Cohen's Kappa
Cohort 1	Occupation and industry	1530	0.367	1542	0.598
	Occupation only	559	0.663	573	0.766
Cohort 2	Occupation and industry	1141	0.489	1189	0.745
	Occupation only	669	0.549	675	0.803

SOC, standard occupational classification.
[a]Number of records coded to six digits by both autocoding programs.
[b]Number of records coded to two digits by both autocoding programs.

choice. Conversely, SOCcer has the ability to process free-text job tasks, so if data contains both job titles and tasks, or have subjects that provide tasks without titles, then SOCcer may be the more appropriate choice. Also, since SOCcer was more effective than NIOCCS when industry was omitted, SOCcer should be used in situations where industry is not available.

Both program performed slightly better in cohort 2 than in cohort 1. A qualitative review showed that cohort 1 contained more ambiguous

TABLE 4 Agreement between O*NET exposures mapped to manually coded SOC codes and that of autocoded SOC codes from pooled upper extremity study (Cohort 1), United States, 2001–2010, and from the show-ME study (Cohort 2), Missouri, 2012–2013

						Intra-class correlations					
Data source	Program	Confidence score	N	O*NET n	% in O*NET	Static strength	Dynamic strength	Handle objects	Repetitive motions	Importance of using computers	General physical activity
Cohort 1	SOCcer	All	1823	1371	75.21	0.712	0.713	0.724	0.464	0.662	0.684
		Less than 0.2	706	551	78.05	0.480	0.462	0.521	0.204	0.372	0.434
		0.2 to 0.4	428	336	78.50	0.809	0.806	0.789	0.548	0.771	0.717
		Greater than .4	689	484	70.25	0.866	0.894	0.854	0.695	0.844	0.869
		90-100 in NIOCCS	1530	1151	75.23	0.716	0.718	0.739	0.480	0.665	0.698
	NIOCCS	All	1530	996	65.10	0.811	0.789	0.842	0.626	0.800	0.733
		100	856	638	74.53	0.869	0.866	0.899	0.711	0.882	0.814
		Less than 100	674	358	53.12	0.718	0.663	0.755	0.499	0.670	0.578
Cohort 2	SOCcer	All	1496	1298	86.76	0.778	0.774	0.709	0.701	0.766	0.706
		Less than 0.2	507	439	86.59	0.612	0.602	0.504	0.542	0.600	0.527
		0.2 to 0.4	476	409	85.92	0.816	0.823	0.760	0.700	0.761	0.735
		Greater than .4	513	450	87.72	0.913	0.904	0.882	0.819	0.915	0.871
		90-100 in NIOCCS	1141	1001	87.730	0.829	0.817	0.764	0.746	0.825	0.778
	NIOCCS	All	1141	941	82.47	0.884	0.886	0.852	0.822	0.901	0.849
		100	756	648	85.71	0.935	0.938	0.920	0.884	0.943	0.912
		Less than 100	385	293	76.10	0.773	0.772	0.741	0.664	0.811	0.740

NIOCCS, NIOSH industry and occupation computerized coding system; O*NET, occupational information network; SOCcer, standardized occupation coding for computer-assisted epidemiological research.

free-text entries than cohort 2. Taking measures to avoid ambiguous occupation text entries should be considered during data collection.

SOCcer lists 10 SOC codes for every data entry providing additional flexibility for the user. However, users are cautioned to review the confidence scores. Even though SOC codes are provided for every entry, a low score may indicate that the associated SOC code is not reliable and should not be used. A priori decisions should be made regarding procedures for using SOCcer output including determining a lowest acceptable confidence score. Also, it must be decided whether to use only the SOC code with highest score, or to consider all the SOC codes with scores that meet the specified cutoff. In practice, it may be beneficial to manually code a random subset of data to get a sense of how well SOCcer performs at various confidence scores. Further investigation in this area may help inform decision making for SOCcer users.

Both autocoding programs may leave a subset of data without adequate SOC codes. In these instances additional manual coding may be necessary. Both programs provide an interface that assists users to manually code any non-coded entries, and to review codes with low confidence scores (or flagged for suggested review in NIOCCS). However, in some circumstances any amount of manual coding may be infeasible or undesirable. Researchers may choose to exclude data that are not supplied adequate codes from a coding program, however, caution should be taken as this may lead to data that are missing not at random. That is, there may be some meaningful differences between subjects that are successfully assigned SOC codes versus those that are not. It should also be noted that availability of O*NET exposure data influenced the manual coding process in both cohorts of the present study. This likely caused some attenuation in agreement since availability of O*NET exposures is not considered by the autocoding programs. Also, since data collection methods, job diversity, and types of job descriptors all differ among datasets, the results of the current study may not be typical.

The use of JEMs in epidemiology is growing. Primary data collection can be very expensive, thus, the ability to combine information from JEMs with pre-existing data is often attractive. Furthermore, JEMs allow the ability to obtain retrospective exposure data, whereas, it is impossible to directly observe such exposures, and self-reported exposures from past jobs may suffer from recall bias or random misclassification. The ability to convert free-text job titles into SOC codes with autocoding programs provides an efficient process for using JEMs.

SOCcer and NIOCCS are powerful tools for assigning SOC codes to free-text occupation entries. This helps facilitate a key step in epidemiological studies of occupational related risk factors. Manual coding can be very time consuming and often times infeasible for large studies. Furthermore, manual coding is an inherently less consistent method than using autocoding programs. Therefore, as the use of SOC codes in research continues to grow, autocoding will likely become standard practice. With this eminent shift, further testing of these autocoding programs is urged.

AUTHORS' CONTRIBUTIONS

SB-P designed and executed the analytic strategy for this study. He was also the primary writer of the manuscript. AMD conducted the manual job coding and assisted with writing. BAE provided guidance, assisted with writing, is the PI on the grant that funded this project, and is the PI on the grant that funded one of the two cohort studies that provided source data. All authors approved the final draft of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank the following groups and individuals whose contributions were necessary for this research: NIOSH Division of Surveillance, Hazard Evaluations, and Field Studies for providing guidance and troubleshooting for use of NIOCCS, Dr Daniel Russ from the NIH for providing guidance and troubleshooting for use of SOCcer, and investigators from the SHOW-ME study (U54 CA155496) and the NIOSH Upper Extremity Musculoskeletal Disorders Consortium (R01OH009712) for access to data from these two studies.

FUNDING

Grant sponsor: National Institute for Occupational Safety and Health; Grant number: R01OH011076.

ETHICS APPROVAL AND INFORMED CONSENT

The analysis was completed using secondary data, which was de-identified; IRB review was not required.

DISCLOSURE (AUTHORS)

The authors declare no conflicts of interest.

DISCLOSURE BY AJIM EDITOR OF RECORD

Steven B. Markowitz declares that he has no competing or conflicts of interest in the review and publication decision regarding this article.

DISCLAIMER

None.

ORCID

Skye Buckner-Petty http://orcid.org/0000-0003-1621-2334

REFERENCES

1. Bureau of Labor Statistics. Average hours per day spent in selected activities on days worked by employment status and sex, 2017 annual averages. In. Washington, DC: U.S. Bureau of Labor Statistics; 2017.
2. Bureau of Labor Statistics. Standard Occupational Classification. 2018; https://www.bls.gov/soc/
3. Dale AM, Ekenga CC, Buckner-Petty S, et al. Incident CTS in a large pooled cohort study: associations obtained by a Job Exposure Matrix

versus associations obtained from observed exposures. *Occup Environ Med*. 2018;75:501–506.
4. Mediouni Z, Bodin J, Dale AM, et al. Carpal tunnel syndrome and computer exposure at work in two large complementary cohorts. *BMJ Open*. 2015;5:e008156.
5. Solovieva S, Pehkonen I, Kausto J, et al. Development and validation of a job exposure matrix for physical risk factors in low back pain. *PLoS ONE*. 2012;7:e48680.
6. Rubak TS, Svendsen SW, Soballe K, Frost P. Total hip replacement due to primary osteoarthritis in relation to cumulative occupational exposures and lifestyle factors: a nationwide nested case-control study. *Arthritis Care Res*. 2014;66:1496–1505.
7. Svendsen SW, Johnsen B, Fuglsang-Frederiksen A, Frost P. Ulnar neuropathy and ulnar neuropathy-like symptoms in relation to biomechanical exposures assessed by a job exposure matrix: a triple case-referent study. *Occup Environ Med*. 2012;69:773–780.
8. Occupational Information Network, U.S. Department of Labor/Employment and Training Administration. O*NET OnLine website. 2018; http://www.onetonline.org/
9. Evanoff B, Zeringue A, Franzblau A, Dale AM. Using job-title-based physical exposures from O*NET in an epidemiological study of carpal tunnel syndrome. *Hum Factors*. 2014;56:166–177.
10. Dale A, Zerinque A, Harris-Adamson C, et al. General population job exposure matrix applied to a pooled study of prevalent Carpal Tunnel syndrome. *Am J Epidemiol*. 2015;181:431–439.
11. Dembe AE, Yao X, Wickizer TM, Shoben AB, Dong XS. Using O*NET to estimate the association between work exposures and chronic diseases. *Am J Ind Med*. 2014;57:1022–1031.
12. Schmitz M, Forst L. Industry and occupation in the electronic health record: an investigation of the National Institute for Occupational Safety and Health Industry and Occupation Computerized Coding System. *JMIR Med Inform*. 2016;4:e5.
13. Weiss N, Cooper S, Socias C, Weiss R, Chen V. Coding of central cancer registry industry and occupation information: the Texas and Louisiana experiences. *J Registry Manage*. 2015;42:103–110.
14. Dodd KE, Mazurek JM. Asthma among employed adults, by industry and occupation—21 states, 2013. *MMWR Morb Mortal Wkly Rep*. 2016;65:1325–1331.
15. O'Halloran AC, Lu PJ, Williams WW, et al. Influenza vaccination among workers-21 U.S. states, 2013. *Am J Infect Control*. 2017;45: 410–416.
16. Russ DE, Ho K-Y, Colt JS, et al. Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies. *Occup Environ Med*. 2016;73:417–424.
17. Rempel D, Gerr F, Harris-Adamson C, et al. Personal and workplace factors and median nerve function in a pooled study of 2396 US workers. *J Occup Environ Med*. 2015;57:98–104.
18. Dale AM, Harris-Adamson C, Rempel D, et al. Prevalence and incidence of carpal tunnel syndrome in US working populations: pooled analysis of six prospective studies. *Scand J Work Environ Health*. 2013;39:495–505.
19. Dale AM, Enke C, Buckner-Petty S, et al. Availability and use of workplace supports for health promotion among employees of small and large businesses. *Am J Health Promot*. 2018; https://doi.org/10.1177/0890117118772510.
20. ReferenceUSA. 2017. http://www.referenceusa.com/Home/Home. Accessed March 08 2017.
21. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6:284.
22. R: A Language and Environment for Statistical Computing [computer program]. Version 3.3.2. Vienna, Austria: R Foundation for Statistical Computing; 2016.

How to cite this article: Buckner-Petty S, Dale AM, Evanoff BA. Efficiency of autocoding programs for converting job descriptors into standard occupational classification (SOC) codes. *Am J Ind Med*. 2019;62:59–68. https://doi.org/10.1002/ajim.22928