Published in final edited form as:

Occup Environ Med. 2016 June ; 73(6): 417–424. doi:10.1136/oemed-2015-103152.

Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies

Daniel E. Russ¹, Kwan-Yuet Ho¹, Joanne S. Colt², Karla R. Armenti³, Dalsu Baris², Wong-Ho Chow⁴, Faith Davis⁵, Alison Johnson⁶, Mark P. Purdue², Margaret R. Karagas⁷, Kendra Schwartz⁸, Molly Schwenn⁹, Debra T. Silverman², Calvin A. Johnson¹, and Melissa C. Friesen²

¹Division of Computational Bioscience, Center for Information Technology, NIH, Bethesda, MD, USA

²Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA

³New Hampshire Department of Health and Human Services, Division of Public Health Services, Bureau of Public Health Statistics and Informatics, Concord, New Hampshire, USA

⁴University of Texas MD Anderson Cancer Center, Houston, Texas, USA

⁵Department of Public Health Sciences, University of Alberta, Edmonton, Alberta, Canada

⁶Vermont Cancer Registry, Burlington, VT, USA

⁷Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA

⁸Department of Family Medicine and Public Health Sciences, Wayne State University, Detroit, MI, USA

⁹Maine Cancer Registry, Augusta, ME, USA

Abstract

Background—Mapping job titles to standardized occupation classification (SOC) codes is an important step in identifying occupational risk factors in epidemiologic studies. Because manual coding is time-consuming and has moderate reliability, we developed an algorithm called SOCcer (Standardized Occupation Coding for Computer-assisted Epidemiologic Research) to assign SOC-2010 codes based on free-text job description components.

Methods—Job title and task-based classifiers were developed by comparing job descriptions to multiple sources linking job and task descriptions to SOC codes. An industry-based classifier was developed based on the SOC prevalence within an industry. These classifiers were used in a logistic model trained using 14,983 jobs with expert-assigned SOC codes to obtain empirical weights for an algorithm that scored each SOC/job description. We assigned the highest scoring SOC code to each job. SOCcer was validated in two occupational data sources by comparing SOC

codes obtained from SOCcer to expert assigned SOC codes and lead exposure estimates obtained by linking SOC codes to a job-exposure matrix.

Results—For 11,991 case-control study jobs, SOCcer-assigned codes agreed with 44.5% and 76.3% of manually assigned codes at the 6- and 2-digit level, respectively. Agreement increased with the score, providing a mechanism to identify assignments needing review. Good agreement was observed between lead estimates based on SOCcer and manual SOC assignments (kappa: 0.6–0.8). Poorer performance was observed for inspection job descriptions, which included abbreviations and worksite-specific terminology.

Conclusions—Although some manual coding will remain necessary, using SOCcer may improve the efficiency of incorporating occupation into large-scale epidemiologic studies.

INTRODUCTION

There is an increasing number of large-scale epidemiologic studies. An on-going challenge in these studies is the processing and use of occupational information to identify occupational risk factors. Crucial occupational information is usually provided by the study subject as free-text responses to open-ended questions, including “What was your job title?” and “What were your main tasks and activities in this job?”, which have been found to be reliably reported.¹² Free-text occupational information is also found in administrative databases, including medical records, registry data, death certifications, and workers’ compensation data.

To use the occupational information for epidemiologic analyses, an important first step is the coding of the study subjects’ responses to standardized occupation classification codes. The codes can be used as an exposure proxy by themselves or as the basis for other ordinal or quantitative exposure assessment methods. For instance, these occupation classification codes can be linked to job-exposure matrices (JEMs) to assign job-level exposure estimates. The coding is usually done manually and – because the coding has only moderate reliability¹³ – it is preferably conducted by multiple independent coders who then resolve discordant assignments. In large studies, however, manual and consensus coding may be cost prohibitive. Coding challenges also occur when pooling studies, which are often coded using different systems since classification systems differ by country and change over time.³⁴

The availability of natural language processing methods has led to the development of computer algorithms to automate coding. Existing algorithms have moderate to moderately-high agreement between the manual and automated assignments.^{5–8} Freely available web tools to aid classifying jobs have also been developed (e.g., <http://wwwn.cdc.gov/niosh-nioccs>; www.caps-canada.ca). These algorithms differ in the classification system used and in their matching processes. For example, Burstyn et al.⁶ used a hierarchical classification approach based on Naïve Bayes; Osslander et al.⁹ used exact word matching; Caps Canada uses exact word match with suggested spelling corrections and incorporates weights depending on word placement; and the current version of the NIOSH Industry and Occupation Computerized Coding System (NIOCCS) uses a combination of approaches (direct/partial/fuzzy match, n-gram match, hard-coded rules, and phonetics) in a tiered approach. The development of these approaches has been hampered by the quality of the job

descriptions. Job titles are often provided as short phrases and recorded with spelling errors and non-standard abbreviations, which causes direct matching to fail. Additionally, a job title may not be specific enough to differentiate between similar, but distinct classifications (e.g., ‘technician’). Moreover, these systems often use only the job title; in manual coding the task and industry information often helps identify the best occupation classification.

In other classification challenges, better predictive performance has been found using ensemble, rather than single, classification methods, especially when there is diversity among the classifiers being integrated.¹⁰¹¹ These diverse classifiers can be combined using a stacked ensemble approach, which involves using a training data set to empirically fit a function that combines the classifiers to produce a single classification result.¹² Depending on the outcome to be classified, stacked ensemble approaches have used a variety of linear, polynomial, and logistic regression models to integrate the individual classifiers into a single prediction.¹¹¹³ This approach provides an opportunity to combine the information from multiple job components, which may vary in their relative strength of association with the occupation classification, to identify the best occupation classification.

Here, our first aim was to describe SOCcer (Standardized Occupation Coding for Computer-assisted Epidemiologic Research, available at <http://soccer.nci.nih.gov>), which is a stacked ensemble algorithm that we developed to assign U.S. SOC-2010 occupation codes¹⁴ based on information provided by free-text job title, free-text tasks, and previously coded industry classifications. Our second aim was to assess whether the SOCcer algorithm could reliably reproduce expert SOC assignments by comparing the SOC codes assigned using SOCcer to expert-assigned SOC codes using two sources of job descriptions with varying qualities of occupational information. Our aim was to provide insights into SOCcer’s performance and to identify where expert review of the algorithm-assigned codes may be necessary.

METHODS

Sources of job descriptions linked to SOC codes

We first describe the data sources that were available to us to assist with the development and evaluation of a stacked ensemble classifier algorithm to code free-text job descriptions because this data provides the algorithm’s foundation and scope.

Publicly available occupational information—We used multiple sources of public data to create a knowledge base that linked job titles and tasks to SOC-2010 codes. Our primary source for both job title and task information was the Occupational Information Network (O*NET) (<http://www.onetonline.org>). Additional job title-SOC code links were added from the Bureau of Labor Statistics Direct Match Title File¹⁵ and the U.S. Census Occupational Index.¹⁶ After removing redundancies, our knowledge base comprised ~62,000 job titles linked to SOC-2010 codes.

Training Data—The development of a stacked ensemble classifier requires a training data set of job descriptions that are similar to the type of job descriptions to be coded. For algorithm training, we used job descriptions obtained from the New England Bladder Cancer Study (NEBCS), which comprised 2631 participants reporting 14,983 jobs¹⁷. The job

descriptions were obtained from study participants using an interviewer-administered lifetime occupational history questionnaire. The descriptions were free-text responses to questions on job title, employer, products and services provided by their employer, their main job tasks and activities, tools and equipment, chemicals and substances, and dates for each job they held. We re-coded the previously coded SOC-1980 assignments for each job to SOC-2010 codes using cross-walks between SOC-1980 and SOC-2000 and from SOC-2000 to SOC-2010. Multiple matches from the crosswalks were not reviewed to obtain a single code and were considered equally likely assignments because NEBCS was used for model training rather than validation.

Evaluation data set 1: OSHA's Integrated Management Information System—

Our first data set used to evaluate whether SOCcer reliably reproduced expert SOC assignments was jobs reported in the U.S. Occupational Safety and Health Administration's (OSHA) Integrated Management Information System (IMIS). These job descriptions, which were collected as documentation for personal, airborne measurements during OSHA inspections, included both the free-text job titles and the inspected company's industry coded to 1987 Standard Industry Classification (SIC) codes. The job titles consisted of two to three word job descriptions entered by the inspector, often containing plant-specific terminology and locations (e.g., #3 lead pot operator), non-standard abbreviations (e.g. FLVE cleaner b), and misspellings (e.g. laboror). IMIS's ~80,000 lead and cadmium measurements (which were extracted for other purposes) provided over 17,100 unique free-text job descriptions/SIC combinations. From these job descriptions, we randomly selected two subsets for which we obtained SOC-2010 codes from expert coders. A subset of 100 jobs was used to assess feasibility and training and a subset of 1000 jobs was used for evaluating SOCcer. Expert coding was obtained in two stages. First, two coders independently assigned SOC-2010 codes to each job. Second, jobs with discordant assignments were reviewed by the two coders, their supervisor, and a senior industrial hygienist to obtain a consensus SOC assignment.

Evaluation data set 2: US Renal Cell Carcinoma Study—Our second evaluation data set consisted of job descriptions from the US Renal Cell Carcinoma Study (hereafter, US Renal), which comprised 2,048 subjects reporting 11,991 jobs.¹⁸¹⁹ The job descriptions were collected similarly to the occupation information in NEBCS. The job descriptions had been previously double-coded to SOC-1980¹⁴ and SIC-1987, with a supervisor resolving discordant codes.²⁰ We cross-walked the SOC-1980 codes to SOC 2000 and then to SOC-2010. Jobs with multiple SOC-2010 matches were resolved by a single coder with access to both the original job descriptions and the cross-walk matches (69% of the jobs). Jobs with insufficient information with which to assign a SOC code were excluded (n=48).

SOCcer algorithm

SOCcer is a stacked ensemble classifier that integrates the results of multiple classifiers (3 from job title, 1 from industry, and 1 from task) to provide a single score that compares a job description to a SOC 2010 code. By design, the classifiers are not necessarily equivalent in importance. For instance, the industry-based classifier was a dichotomous variable that identifies a set of SOC codes that were frequently observed within that industry and thus

provided multiple plausible SOC matches. To obtain an algorithm that provided a single score for each job-SOC code comparison, we trained the five classifiers' results on previously coded job descriptions using a logistic regression model. The model provided empirical-based weights that scaled the classifiers' results, quantified their relative importance, and provided a score that estimated the probability that an expert coder would have assigned a SOC code to the job description. For each job, 840 SOC-specific sets of classifier results and SOCcer scores were calculated. The SOCcer output provides the 10 highest scoring SOC codes for each job description. Here, jobs in the two evaluation data sets were assigned the SOC code with the highest score, which was then compared to the expert-assigned SOC code. Below we provide a brief description of each of SOCcer's components.

Job title-based classifiers—Prior to developing the classifiers, the job titles from all sources were pre-processed: non-letter characters were replaced with spaces; capital letters were converted to lower-case, and job titles comprising multiple words were tokenized (“assembler packer” became “assembler” and “packer”). The first job title classifier was developed using a similarity measure derived using a *soft Jaccard* index, J , that compared the observed job title to job titles in our knowledge base. J accounted for close matches, such as word variants, misspellings, and abbreviations, by giving partial credit for similarly spelled words²¹, as described in Appendix A. A second job title classifier, J_{max} , was a dichotomous variable that identified the SOC code(s) that had the highest soft Jaccard index for that job title (1 if maximum for that job title; 0 if not maximum). A third job title classifier, P_{maxent} , was a maximum entropy classifier^{22,23} that we incorporated because preliminary results using only J and J_{max} had a large number of ties. P_{maxent} was estimated based on the probability of a job title-SOC code match being observed in the training data and, when that combination was not observed in the training data, a uniform probability was assumed across the SOC codes. P_{maxent} was built using the Apache OpenNLP Maxent toolkit²⁴.

Industry-based classifier—Industry information was incorporated because the SOC assignments in SOC-2010 were often industry-specific. For example, a fireman with SIC code 9224 (Fire Protection) is more likely to be a fire fighter (SOC 33-2011), while a fireman with SIC code 3269 (Pottery Products) is more likely to be a kiln operator (SOC 51-9051). As a result, we developed an industry-based classifier, $S > 0.01$, that equally up-weighted the set of SOC codes that were observed with 1% frequency within an industry. The prevalence of a SOC code within each SIC (0 to 1) was obtained using the National Industry-Occupation Employment Matrix²⁵ from the U.S. Bureau of Labor statistics, after applying a SIC to NAICS crosswalk. The prevalence was dichotomized such that all SOC codes meeting a specified threshold were treated as equally plausible assignments for that industry. Sensitivity analyses with thresholds set from 0.001 through 0.05 were robust, with the best model fit observed with 0.01 (not shown).

Task-based classifier—To compare the job task information to the task information for each SOC code provided in the O*NET database, we built a classifier, FPP , using a fuzzy fingerprint approach that compared each task description to a task fingerprint for each SOC

code (see Appendix A). The fingerprint was a list of associated task words and a measure of each word's importance²⁶.

Aggregating the classifiers—Logistic regression models have been commonly used to aggregate multiple classifiers into a single classification¹². The model is used to estimate the likelihood that the combined classifier results predict the correct classification assignment, while allowing the importance and scaling of each classifier to be empirically estimated based on the training data. Here, the five classifiers (J , J_{max} , P_{maxent} , $S > 0.01$, and FPP) were applied to the 14,983 NEBCS jobs to obtain classifier values for each job-SOC code comparison (14,983 job $\times$ 840 SOC comparisons). For each comparison, the outcome variable identified whether that SOC code had been assigned to that NEBCS job description (yes=1; no=0). The NEBCS classifiers' values were evaluated in a logistic regression model using the glm function of R²⁷ with the structure shown in equation 1, where y was the log-odds of the likelihood that an expert assigned that SOC code to a job description. Non-linear terms for the continuous classifiers were tested but provided negligible improvement in fit (not shown).

$$y = \beta_0 + \beta_1(J) + \beta_2(J_{max}) + \beta_3(P_{maxent}) + \beta_4(FPP) + \beta_5(S > 0.01) \quad (1)$$

This model provided empirically-based weights that were used with the classifiers' results (equation 1) to obtain a score for each job-SOC code comparison (equation 2). The score was interpreted as the estimated probability that an expert coder would have assigned that SOC to the job description. When a component was lacking in the job description (e.g., no industry information available), the associated classifier value was set to 0. The SOCcer output provides the 10 highest scoring SOC codes for each job description.

$$\text{Score} = \frac{e^y}{e^y + 1} \quad (2)$$

Figure 1 shows an example job description for a plumber who repairs pipe and who works in the industry 'General Contractors-Single-Family Houses' (SIC-1987: 1521). In the example, the algorithm selects SOC-2010 47-2152 (Plumbers, Pipefitters, and Steam fitters) as the SOC code with the highest SOCcer score.

SOCcer evaluation

Exact matches to expert-assigned SOC codes—For the jobs in the two evaluation sets, we applied the five classifiers and then calculated the SOCcer score for each job-SOC code comparison using equations 1 and 2, using the β 's obtained from training the model on the NEBCS jobs. The SOC code with the highest SOCcer score was assigned to the job description. For each source, we calculated the proportion of jobs to which SOCcer assigned multiple SOC codes because of ties in the highest score (exact match at 6th decimal place). Ties were considered mismatches because the assignment required expert review. We also calculated the proportion of jobs where the SOCcer and the expert's SOC code matched at the 2-, 3-, 5-, and 6-digit level. Jobs where a coder provided only a lower detail level (e.g., 3- or 5-digit) were excluded from the comparisons at more detailed levels (e.g., 6-digit).

Agreement between JEM-based lead estimates—Discrepancies in SOC codes do not necessarily impact exposure estimates, especially for administrative and management jobs. To evaluate whether the SOCcer and expert's SOC assignments would have resulted in the same exposure estimates, we applied a lead JEM linking the SOCcer and manually-assigned SOC codes of the IMIS and US Renal jobs to ordinal (0 to 3) estimates of the probability and intensity of occupational lead exposure. Lead was chosen as the first agent-specific evaluation for convenience purposes because of on-going lead exposure assessment efforts within the study group. The lead JEM had been previously developed for other populations^{28–30} and was cross-walked to SOC-2010 manually by an industrial hygienist for these analyses. Agreement between exposure estimates was calculated using Cohen's weighted kappa with quadratic weights using all the available job components in each case study, both overall and by SOCcer score. To examine the contribution of each job component to the algorithm's performance, we also calculated the agreement measures for SOCcer assignments that were based on job title alone, job title and task only, and job title and industry only.

Comparison to NIOCCS—For the US Renal job descriptions, we also compared SOCcer's performance to NIOCCS's performance, which is a potential alternative for automatically coding jobs to SOC-2010. We did not evaluate the IMIS jobs from this evaluation because of the poorer quality of the job descriptions. We applied NIOCCS to the US Renal jobs and obtained NIOCCS-assigned SOC-2010 codes when the NIOCCS system estimated the assignment as having a 'medium' or 'high' confidence; the definitions of these confidence assessments were not provided.⁵ We compared the agreement between the SOC codes assigned by NIOCCS to those obtained from an expert coder and from SOCcer.

RESULTS

Overall agreement: IMIS

For IMIS data, the two coders assigned the same SOC code for 50.4%, 56.8%, 68.6%, and 86.2% of the jobs based on the 6-, 5-, 3-, and 2-digit SOC levels, respectively. Comparisons of lead exposure estimates obtained using the two coders' 6-digit SOC assignments observed weighted kappas of 0.72 for lead probability and 0.65 for lead intensity.

Table 1 shows the comparisons of SOCcer- and expert consensus-assigned SOC codes for the IMIS data. Using SOCcer, two or more SOC codes tied for highest score for 4.6% of the jobs using all available job description components and 3.8% of the jobs using only the job title component. The highest scoring SOC code from SOCcer using both job title and industry matched the expert consensus SOC code for 36.8% and 70.4% of the jobs at the 6-digit and 2-digit levels, respectively. JEM-based lead estimates obtained using the highest scoring SOC code from SOCcer and the consensus codes had moderate agreement ($\kappa=0.49$ for intensity, 0.45 for probability). Exact matches decreased slightly when only job title was used (35.1% agreement at 6-digit SOC; $\kappa=0.36$, 0.45).

Overall agreement: US Renal

Agreement between SOCcer- and expert-assigned SOC codes for US Renal jobs are shown in Table 1, with higher agreement observed in US Renal jobs than in IMIS jobs. Using all job components and classifiers, two or more SOC codes tied for highest SOCcer score for 1.1% of the US Renal jobs. The SOCcer- and expert-assigned SOC codes matched for 44.5% and 76.3% of the US Renal jobs at the 6- and 2-digit level, respectively. High agreement was observed between JEM-based lead exposure estimates obtained using the expert-assigned SOC code and the highest scoring SOC code from SOCcer ($\kappa=0.78$ for probability, 0.74 for intensity). Compared to using all job components, SOCcer assignments made when either the task and/or industry job description component was excluded had lower agreement by ~1–3% at each SOC hierarchical level; however, the exclusion of these components had negligible impact on agreement between lead exposure estimates obtained using expert vs. SOCcer-assigned SOC codes.

Agreement by SOCcer score

Figure 2 shows that IMIS's median SOCcer score of 0.25 was substantially lower than US Renal's median of 0.47, which was driven by IMIS's large peak below 0.15. In these two data sets, increasing score was associated with improved agreement between SOCcer and coder(s) (Figure 3) and with improved agreement in lead exposure estimates (Figure 4), which plateaued at scores above 0.2.

We further evaluated the agreement between SOCcer- and expert-assigned SOC codes by occupation group (Supplementary Table S1). Higher agreement in these subgroups generally corresponded to higher median SOCcer scores; similarly, lower agreement corresponded to lower scores. The Spearman correlation coefficient between agreement at the 6-digit level and the median SOCcer score was 0.60 (not shown).

Comparison of SOCcer and NIOCCS

NIOCCS provided assignments for only 3,332 (28%) US Renal jobs; the remaining assignments were not provided by NIOCCS because they had less than a 'medium' level confidence. Overall, considering non-assigned jobs as a mismatch because they would require expert review, the agreement between NIOCCS and coders was 17.3% at the 6-digit level. For the subset of jobs with a NIOCCS assignment, the agreement between NIOCCS and coders was 61.9% and between SOCcer and coders was 59.2%. Within this subset, SOCcer and NIOCCS agreed 70.9% of the time. When the two systems agreed, their agreement with coders was 75.3%. When the two systems disagreed, NIOCCS, SOCcer, or neither agreed with the coder for 29.3%, 20.2%, and 50.5% of the jobs, respectively.

DISCUSSION

We developed an ensemble classifier that automatically assigned SOC codes based on free-text job titles, free-text tasks, and SIC-1987. In external comparisons of jobs obtained from a case-control study, SOCcer's assignments had overall agreement with expert coders that ranged from 45% at the 6-digit level to 76% at the 2-digit level and was similar to the moderate to moderately-high agreement between expert coders found in previous

studies¹³³¹³². Overall agreement was poorer for occupational data reported alongside inspection measurements. However, in both data sets, the agreement improved with increasing SOCcer score. The similar magnitude of the agreement at similar scores suggests that the score provides a metric with which one could prioritize jobs for manual coding.

SOCcer's overall performance varied in the two evaluation data sets. SOCcer's poorer performance in IMIS likely reflected the poorer quality of the job descriptions, which contained non-standard abbreviations and superfluous information within a job title such as assembly line number. Although relevant to the inspection, this information was not available in our knowledge base and thus resulted in poor matches. Given these low scores and the nature of the job descriptions, most IMIS jobs would require expert review. In contrast, US Renal provided occupational information from a typical epidemiologic study and had a much higher median SOCcer score. This suggests that these job descriptions were a better match with the information in the knowledge base and training data and that a smaller proportion would require expert review. In both IMIS and US Renal, the agreement with expert coders was similar when stratified by SOCcer scores, suggesting that the relationship between agreement and score is reasonably independent of the dataset. Similarly, the score accounted for observed differences in agreement by occupation subgroups. As a result, we believe that the score is a useful metric for prioritizing expert review of SOCcer-coded jobs. The score thresholds for expert review would be based on study-specific goals.

SOCcer's overall performance for the US Renal jobs is consistent with the agreement found using other computerized coding systems⁵⁻⁷⁹. For US Renal, SOCcer had ties for only 1% of the jobs and had agreements with coders' SOC codes of 44.5% and 76.3% at the 6- and 2-digit level. Moreover, SOCcer performed similarly to NIOCCS (59.1% vs. 61.9%, respectively) for a subset of US Renal jobs that met the NIOCCS's medium confidence threshold. Unfortunately, 72% of these jobs did not meet the NIOCCS confidence threshold and would have required manual coding. NIOCCS primary aim was to code industry and occupation data to US Census Bureau classification schemes and, although SOC-2010 codes are provided when a medium confidence threshold is crossed, NIOCCS documentation indicated that SOC-2010 codes should be used cautiously as it was not its primary aim. In contrast, SOCcer was designed specifically to code epidemiologic occupation data to SOC-2010 codes and to provide a score for all assignments.

Few automatic coding systems explicitly incorporate task and industry information into their framework. Incorporating these additional components reduced ties by 1% and improved agreement by a modest 1-3%. The improvement in performance was limited because only O*NET task information was used. In future iterations, the knowledge base could be augmented to include task information reported in epidemiologic studies or from resources similar to O*NET to improve the contribution of the task component. The industry component also provided only a slight improvement over using only job title because it was limited to up-weighting the scores of a set of SOC codes frequently observed in that industry. It was included because some administrative data sets may be pre-coded to SIC or similar industry classification systems. However, most epidemiologic studies that require occupation coding will also require industry coding. We designed SOCcer to be flexible

such that it can use all or a subset of the classifiers by setting a missing classifiers' values to 0.

When the SOC assignment is to be used as the basis of an exposure assessment, it may be sufficient to obtain agreement on the exposure estimates rather than the SOC code. For instance, administrative, sales, and management jobs represent a substantial portion of the jobs in the population and distinguishing between nuances in these codes is likely challenging for SOCcer because of their similar terminology; however, these differences may not change their exposure assignment. Here, in evaluations based on a single agent, agreement based on JEM-based lead estimates increased rapidly as SOCcer scores increased to 0.2, with modest increases at scores greater than 0.2. Future evaluations to examine the generalizability of this finding with other agents are needed.

SOCcer's strengths include its flexibility to incorporate multiple components of the job description in an ensemble classifier, its adaptability for future information, and its provision of a SOCcer score for the ten highest scoring SOC codes for each job. Researchers can use the score to prioritize expert review and can define 'ties' either as exact matches or scores within a user-specified distance. In addition, it was evaluated against manually coded SOC codes in two studies with job descriptions that represented better quality (i.e., US Renal) and lesser quality (i.e., IMIS) occupational data to provide a first estimate of SOCcer's performance. Further evaluations are needed to more fully characterize SOCcer's performance.

SOCcer's main challenges were in identifying suitable information sources that linked job description components to SOC codes and the lack of gold standards with which to evaluate SOCcer. For example, we did not use the 'tools and equipment' or 'chemicals and material' questions as we have not yet found any sources that link these to job titles or to SOC codes. We continue to compile multiple epidemiologic studies and other information sources to increase our knowledge base and incorporate additional classifiers. Because no gold standard coding system exists, the performance of computerized SOC coding systems is evaluated against expert coders. This remains an uncertain comparison because agreement between experts varies from 44–89% at the most detailed occupation classification levels¹³³¹³². For the IMIS jobs, our inter-coder agreements of 50% and 57% at the 6- and 5-digit levels, respectively, were consistent with these prior findings. Similarly, the moderately-high between expert agreement for JEM-based lead estimates ($kw=0.65-0.72$) is consistent with previous findings for multiple agents ($kw=0.66-0.84$)³¹. As a result, our evaluation of SOCcer quantifies how well SOCcer reproduces a coder's job-by-job assignment rather than its validity, which cannot easily be assessed. An additional limitation is our use of cross-walks between coding systems which may also introduce misclassification in SOC codes. In addition, SOCcer does not yet address extracting the standardized industry classification codes from the free-text job descriptions.

In summary, our system provides a computer-based method of coding occupational information to SOC codes, which may allow researchers to study occupational risks in large populations that may otherwise be prohibitively expensive to code. SOCcer's current performance in good quality occupational data is consistent with the agreement often

observed between two coders, with the added benefit of providing a score that can be used to prioritize expert review. SOCcer does not completely replace expert coding, but the provided score can be used to reduce the number of jobs that need expert review.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Sue Pan, Robert Shirley, and Carl McCabe (CBIIT, NCI) for their collaboration to make SOCcer available online. This project was funded by the Intramural Research Programs of the Center for Information Technology, NIH, and of the Division of Cancer Epidemiology and Genetics, NCI, NIH.

REFERENCES

1. Kromhout H, Vermeulen R. Application of job-exposure matrices in studies of the general population: some clues to their performance. *Eur Respir Rev*. 2001
2. Teschke K, Olshan AF, Daniels JL, et al. Occupational exposure assessment in case-control studies: opportunities for improvement. *Occupational and environmental medicine*. 2002; 59(9):575–593. discussion 94. [PubMed: 12205230]
3. 't Mannetje A, Kromhout H. The use of occupation and industry classifications in general population studies. *Int J Epidemiol*. 2003; 32(3):419–428. [PubMed: 12777430]
4. Kogevinas M. Commentary: standardized coding of occupational data in epidemiological studies. *Int J Epidemiol*. 2003; 32(3):428–429. [PubMed: 12777431]
5. U.S. Centers for Disease Control and Prevention. NIOSH Industry and Occupation Computerized Coding System (NIOCCS). Secondary NIOSH Industry and Occupation Computerized Coding System (NIOCCS). 2015. <http://www.cdc.gov/niosh-nioccs/>
6. Burstyn I, Slutsky A, Lee DG, et al. Beyond crosswalks: reliability of exposure assessment following automated coding of free-text job descriptions for occupational epidemiology. *Ann Occup Hyg*. 2014; 58(4):482–492. [PubMed: 24504175]
7. Patel MD, Rose KM, Owens CR, et al. Performance of automated and manual coding systems for occupational data: A case study of historical records. *American Journal of Industrial Medicine*. 2012; 55(3):228–231. [PubMed: 22420026]
8. Elias P, Ellison R, Jones R. EurOccupations: CASCOT software for coding job titles 2009. http://www.wageindicator.org/main/copy_of_documents/policy-briefs/European-Policy-Brief-no-3-CASCOT-coding-program-EUROCCUPATIONS-20100104.pdf.
9. Osslander EM, Milham S. A computer system for coding occupation. *Am J Ind Med*. 2006; 49(10): 854–857. [PubMed: 16804909]
10. Feature engineering and classifier ensemble for KDD Cup 2010. 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010); 2011; Washington, DC.
11. Töschler A, Jahrer M, Bell RM. The BigChaos Solution to the Netflix Grand Prize. 2009 http://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf.
12. Wolpert DH. Stacked Generalization. *Neural Networks*. 1992; 5(2):241–259.
13. Wang H, Xu QS, Zhou LF. Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble. *Plos One*. 2015; 10(2)
14. U.S. Bureau of Labor Statistics. Standard Occupational Classification System (SOC) System. Secondary Standard Occupational Classification System (SOC) System. 2010. <http://www.bls.gov/soc>
15. U.S. Bureau of Labor Statistics. Direct Match Title File. Secondary Direct Match Title File. 2010. <http://www.bls.gov/soc/materials.htm>

16. U.S. Census Bureau. Industry and Occupation – Indexes – People and Households – US Census Bureau. Secondary Industry and Occupation – Indexes – People and Households – U.S. Census Bureau. 2014. <http://www.census.gov/people/io/methodology/indexes.html>
17. Colt JS, Karagas MR, Schwenn M, et al. Occupation and bladder cancer in a population-based case-control study in Northern New England. *Occupational and environmental medicine*. 2011; 68(4):239–249. [PubMed: 20864470]
18. Colt JS, Schwartz K, Graubard BI, et al. Hypertension and risk of renal cell carcinoma among white and black Americans. *Epidemiology*. 2011; 22(6):797–804. [PubMed: 21881515]
19. Purdue MP, Colt JS, Graubard B, et al. A case-control study of reproductive factors and renal cell carcinoma among black and white women in the United States. *Cancer Causes Control*. 2011; 22(11):1537–1544. [PubMed: 21866373]
20. United States. Standard industrial classification manual : 1987. Washington, D.C.: The Office : Supt. of Docs.; 1987. Office of Management and Budget.
21. Russ DE, Ho KY, Johnson CA, et al. Computer-Based Coding of Occupation Codes for Epidemiological Analyses. *Proceedings / IEEE International Symposium on Computer-Based Medical Systems IEEE International Symposium on Computer-Based Medical Systems*. 2014; 2014:347–350. [PubMed: 25221787]
22. Berger AL, DellaPietra SA, DellaPietra VJ. A maximum entropy approach to natural language processing. *Computational Linguistics*. 1996; 22(1):39–71.
23. Manning, CD.; Schtze, H. Foundations of statistical natural language processing. Cambridge, Mass: MIT Press; 1999.
24. The Apache Software Foundation. Apache OpenNLP. Secondary Apache OpenNLP. 2010 <https://opennlp.apache.org/>.
25. U.S. Bureau of Labor Statistics. Industry-Occupation Employment Matrix. Secondary Industry-Occupation Employment Matrix. <http://data.bls.gov/oep/nioem/empiohm.jsp>
26. Authorship identification and author fuzzy "fingerprints". *Fuzzy Information Processing Society (NAFIPS)*, 2011 Annual Meeting of the North American; 18–20 March 2011; 2011.
27. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: 2014.
28. Cocco P, Dosemeci M, Heineman EF. Brain cancer and occupational exposure to lead. *J Occup Environ Med*. 1998; 40(11):937–942. [PubMed: 9830598]
29. Cocco P, Ward MH, Dosemeci M. Occupational risk factors for cancer of the gastric cardia. Analysis of death certificates from 24 US states. *J Occup Environ Med*. 1998; 40(10):855–861. [PubMed: 9800169]
30. Cocco P, Heineman EF, Dosemeci M. Occupational risk factors for cancer of the central nervous system (CNS) among US women. *Am J Ind Med*. 1999; 36(1):70–74. [PubMed: 10361589]
31. Koeman T, Offermans NS, Christopher-de Vries Y, et al. JEMs and incompatible occupational coding systems: effect of manual and automatic recoding of job codes on exposure assignment. *Ann Occup Hyg*. 2013; 57(1):107–114. [PubMed: 22805748]
32. MacDonald LA, Pulley L, Hein MJ, et al. Methods and feasibility of collecting occupational data for a large population-based cohort study in the United States: the reasons for geographic and racial differences in stroke study. *BMC Public Health*. 2014; 14:142. [PubMed: 24512119]

WHAT THIS PAPER ADDS

- In epidemiology studies, coding job descriptions to a standardized occupational classification system is often an important first step in evaluating occupational risk factors, but it is time-consuming and expensive.
- We developed a classification algorithm to automatically code free-text job descriptions using machine learning techniques based on natural language processing methods.
- Evaluations of this automated classification algorithm in two case studies found that it reliably replicated manually assigned occupation codes. Its agreement with manual coding increased with algorithm score, which can be used to identify those job descriptions needing manual review.
- A web-based implementation of the algorithm, which is available at <http://SOCcer.nci.nih.gov>, can assist epidemiological researchers in incorporating occupational risk factors into their studies.

$$y = \ln(\text{odds}) = \ln\left(\frac{p}{1-p}\right) = f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

$$\text{score} = \frac{e^y}{e^y + 1}$$

Job Title	Tasks	SIC	SOC	x_1	x_2	x_3	x_4	x_5	Score
Plumber	Repair Pipes	1521	11-1011	0.00	0	4.70×10^{-4}	0.00	0.00	0.007
Plumber	Repair Pipes	1521	11-1021	0.00	0	4.70×10^{-4}	0.00	0.00	0.007
...	...	...	...	...	...	...	...	...	...
*Plumber	Repair Pipes	1521	47-2152	1.00	1	0.267	1.00	0.039	0.827
...	...	...	...	...	...	...	...	...	...
Plumber	Repair Pipes	1521	53-7199	0.56	0	4.70×10^{-4}	0.00	0.00	0.013

*Highest Score for data row

Figure 1.

An overview of the model used to assign SOC codes to job titles. The input consists of the free-text job title, free-text tasks, and industry previously coded to SIC-1987. For each job description and SOC code combination, the values of five classifiers are calculated. The highest Jaccard score (J) and the SOC prevalence ($S > 0.01$) are dichotomous variables, and the rest are continuous. In the example, Job Title = "Plumber", Tasks = "Repair Pipes", and SIC = 1521 (General Contractors – Single Family Homes), scores are calculated for every SOC. The scores are listed for the following SOC codes: 11-1011, Chief Executives; 11-1021, General and Operations Managers; 47-2152, Plumbers, Pipefitters, and Steamfitters; 53-7199, Material Moving Workers, All Others. The SOC code with the highest score, 47-2152, would be assigned to this job.

Figure 2.

The distribution of the highest SOCcer score for the jobs in the two evaluation sets shows differences between the IMIS (triangles) and US Renal (crosses): the IMIS data has a large number of jobs with very low scores and its median score (0.25) is lower than the US Renal data's median score (0.47), which reflect the underlying quality of the free-text responses.

Figure 3.

Agreement between SOCcer and the consensus SOC codes as a function of the highest observed SOCcer score for each job in the two evaluation data sets.

Figure 4. Agreement between SOCcer and coder assigned SOC codes for the IMIS (triangles) and US Renal (crosses) data. The agreement rapidly increases for very low SOCcer scores, but quickly reaches a maximum value.

Comparisons between the expert-assigned SOC code and the SOC code with the highest SOCcer score at varying hierarchical SOC code levels and for varying combinations of classifiers that incorporated different job description components (Equations 1 and 2). Unused classifiers were assigned a value of 0.

Table 1

	% highest SOCcer score ties ^a	Proportion of jobs with exact match between expert and SOCcer assigned SOC code				Weighted kappa between lead estimates obtained using expert vs. SOCcer-assigned SOC codes	
		2-digit	3-digit	5-digit	6-digit	κ_p	κ_i
<u>IMIS</u> ^b							
Job Title + Industry ^c	4.6%	70.4%	48.8%	38.6%	36.8%	0.45	0.49
Job Title	3.8%	67.1%	44.9%	36.8%	35.1%	0.36	0.45
<u>US Renal</u>							
Job Title + Industry + Task ^c	1.1%	76.3%	63.8%	51.5%	44.5%	0.78	0.74
Job Title + Industry	2.5%	74.8%	62.7%	49.8%	43.4%	0.78	0.75
Job Title + Task	1.2%	74.9%	61.6%	50.3%	43.4%	0.78	0.73
Job Title	2.7%	73.2%	60.6%	48.8%	42.5%	0.77	0.73

κ_p , weighted kappa for the lead probability estimates; κ_I , weighted kappa for the lead intensity estimate.

^aProportion of jobs where two or more SOC codes tied for highest SOCcer score. Because a single code was not assigned, these jobs were treated as discordant with the expert-assigned SOC code even if one of the tied codes matched the expert-assigned code.

^bSOCcer SOC assignment compared to consensus expert SOC code.

^cFull model using all available job components available for that case study.