Published in final edited form as:

Am J Ind Med. 2012 March ; 55(3): 228–231.

Performance of Automated and Manual Coding Systems for Occupational Data: A Case Study of Historical Records

Mehul D. Patel, MSPH[1], **Kathryn M. Rose, PhD**[1], **Cindy R. Owens, MPH**[1], **Heejung Bang, PhD**[2], and **Jay S. Kaufman, PhD**[3]

[1]Department of Epidemiology, University of North Carolina at Chapel Hill

[2]Department of Public Health Sciences, University of California at Davis

[3]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University

Abstract

Background—Occupational data are a common source of workplace exposure and socioeconomic information in epidemiologic research. We compared the performance of two occupation coding methods, an automated software and a manual coder, using occupation and industry titles from U.S. historical records.

Methods—We collected parental occupational data from 1920–40's birth certificates, Census records, and city directories on 3,135 deceased individuals in the Atherosclerosis Risk in Communities (ARIC) study. Unique occupation-industry narratives were assigned codes by a manual coder and the Standardized Occupation and Industry Coding software program. We calculated agreement between coding methods of classification into major Census occupational groups.

Results—Automated coding software assigned codes to 71% of occupations and 76% of industries. Of this subset coded by software, 73% of occupation codes and 69% of industry codes matched between automated and manual coding. For major occupational groups, agreement improved to 89% (kappa=0.86).

Conclusions—Automated occupational coding is a cost-efficient alternative to manual coding. However, some manual coding is required to code incomplete information. We found substantial variability between coders in the assignment of occupations although not as large for major groups.

Keywords

Occupational Coding; Automatic Data Processing; Computer Systems; Occupation Classification; Social Class

Introduction

Occupational and industry titles from contemporary and historical records are common sources of workplace exposures and socioeconomic data in epidemiologic research. However, these data, typically in narrative form, must be translated into meaningful titles or codes. Historically, trained coders manually assigned industry and occupation codes based

Corresponding author: Mr. Mehul D. Patel, MSPH, Cardiovascular Disease Program, 137 E. Franklin Street, Suite 306, Chapel Hill, NC 27514. mdpatel@email.unc.edu, Phone: (919) 966-1967, Fax: (919) 966-9800.

Conflict of Interest Statement: We have no conflicts of interest to declare

on a standardized classification system, such as the U.S. Bureau of the Census (BOC) Alphabetical Index of Industries and Occupations [U.S. Bureau of Census, 1992]. This method is labor-intensive, and currently, there is limited availability of trained coders. Further, there are few published studies on the variability between and within coding methods [Mannetje and Kromhout, 2003; Kogevinas, 2003].

Automated computer coding programs are alternatives to manual coding [Mannetje and Kromhout, 2003; Kogevinas, 2003; Bushnell, 1997; Ossiander and Milham, 2006]. However, their performance has not been sufficiently studied. The National Institute for Occupational Safety and Health (NIOSH) has developed the Standardized Occupation and Industry Coding (SOIC) system - a free software package that reads occupation and industry narratives and assigns 3-digit numerical occupation and industry codes based on the 1990 BOC Alphabetical Index of Industries and Occupations [NIOSH, 2001]. We compared the performance of the SOIC automated coding software to that of a manual coder using occupations and industries from historical records. We further assessed the repeatability of manual coding. Finally, we offer recommendations for the integration of these approaches in epidemiologic research.

Methods

As part of research focused on acquiring early life socioeconomic data for decedents from the Atherosclerosis Risk in Communities (ARIC) study [The ARIC Investigators, 1989; Rose et al., 2008] we collected parental occupational data for 3,135 deceased individuals from three sources: 1922–1945 birth certificates, 1930 declassified Census records, and city directories (1924–1949).

From these sources, we identified 2,454 unique occupation–industry narratives for fathers. These were sent, in two batches 15 months apart, to an experienced NIOSH-trained coder for assignment of codes based on the 1990 BOC Alphabetical Index of Industries and Occupations.

The same occupation and industry titles were also run through the SOIC software. For 47 records, the manual coder could not assign an industry code since an unfamiliar business or company name was provided in that field. Therefore, we provided the city and state for these select records to assist in coding. However, no additional information can be processed by the SOIC for coding; thus, none was provided. To assess repeatability of the manual coder, in the second batch, we also included a blinded, 10% random sample of 165 occupation-industry narratives from the first submission. The occupation codes assigned by the manual and SOIC coders were categorized into the major Census occupational groups (Table I). We further dichotomized these into manual (e.g. laborers, craftsmen, farmers) and non-manual (e.g. managers, salesmen) occupations, categories typically used to assign occupation-based socioeconomic status (SES).

In our comparison of automated coding to manual coding, as well as the repeatability of codes assigned by the manual coder on separate submissions, percent agreement was calculated between the following: 3-digit occupational and industry codes; major Census occupational groups; and dichotomous categorization of manual and non-manual occupations. We calculated simple kappa statistics [Fleiss et al., 2003] and 95% confidence intervals for the latter two comparisons in SAS 9.2 (SAS, Cary, NC).

Results

While the expert coder manually assigned 3-digit Census codes to all 2,454 occupation and industry narrative combinations, the SOIC assigned 1,737 (71%) occupation and 1,862

(76%) industry codes. For the records assigned codes by both SOIC and the manual coder, 1,262 (73%) occupation and 1,283 (69%) industry codes matched, and 814 (56%) had both occupation and industry codes match. Table 1 shows the distribution of occupations assigned by each coder when categorized into the standard Census occupation groups. Both coders assigned the majority of occupations into the “Operators, Fabricators, and Laborers” and “Precision Production, Craft, and Repair occupations” groups. Further, of the 717 occupations not coded by the automated system, slightly over half (52%) were manually coded as “Operators, Fabricators, and Laborers”.

For the major occupational groups, manual and automated coding matched on 1,550 of 1,737 (89%) records, [kappa (95% CI) = 0.86 (0.85, 0.88)]. One-third of the discordance was either between the “Precision Production, Craft, and Repair occupations” and “Operators, Fabricators, and Laborers” groups or the “Managerial and Professional Specialty occupations” and “Technical, Sales, and Administrative Support occupations” groups. Agreement improved when occupations were further grouped into manual and non-manual categories [96%; kappa (95% CI) = 0.90 (0.88, 0.92)].

In the repeatability substudy of 165 records coded manually, 125 (76%) occupation, 142 (86%) industry, and 117 industry and occupation (71%) codes matched across the two submissions. When categorized into major occupational groups, agreement was 91% [kappa (95% CI) = 0.87 (0.81, 0.93)]. Agreement was 95% [kappa (95% CI) = 0.87 (0.80, 0.95)] when occupations were further categorized into manual and non-manual occupations.

Discussion

The SOIC computer program left a substantial proportion of historical records-based occupation and industry titles uncoded (29% and 24%, respectively). By design, the SOIC does not assign codes to ambiguous or incomplete information. Furthermore, text entry errors, misspellings, transposed occupation and industry, and multiple occupations and industries listed are a few of the reasons why an automated coder could not code all records. Therefore, in most studies, some manual coding will be needed, depending on the number of records and data quality.

Our findings are consistent with an evaluation of the SOIC conducted by the NIOSH, in which their software agreed with an expert manual coder on 75% of occupation codes and 76% of industry codes [NIOSH, 2001]. In our study, the relatively poor agreement of occupation and industry codes between manual and automated coder (73% and 69%, respectively) suggests considerable subjectivity in the coding process, even in trained, experienced professionals. While an automated system codes consistently on repeated submissions, manual coding is also subject to intra-coder variability. Similar to the comparisons between the manual and automated coders, we found agreement within the manual coder to be moderate for occupation and industry codes but greatly improved with the major occupational groups and manual/non-manual occupation.

Our study underscores the importance of including reliability assessments in coding procedures, particularly when using specific occupational titles to assign workplace exposures. Studies using occupation to measure SES are typically interested in broad groupings, in which case agreement is very good.

There are major limitations in our study. First, we used an occupation and industry classification system developed for the 1990 Census to code historical records from at least 50 to 60 years ago. However, we were restricted to this coding system since this aspect of the SOIC program cannot be modified. Second, we were unable to assess the accuracy of codes due to lack of a gold standard coding method, as was the case with the NIOSH

evaluation. Third, the manual coder was provided with additional city and state information to assist in coding select records because the main purpose of coding was to determine early life SES with the information available. However, this makes comparability to the SOIC, which cannot process additional information, less than ideal, and furthermore, our manual coding procedures may not be reproducible in situations where additional information is not available. Still, given the relatively small number of records manually coded using city and state (2%), these concerns are minimal. Finally, although we quantified variability within a manual coder, we did not assess the additional source of variability between two different manual coders. Nonetheless, this study provides novel findings pertinent to an important potential source of error.

In conclusion, we recommend investigators choose their coding approach based on the study size and intended use of occupational data (e.g. workplace exposure, SES). Automated coding will typically need to be supplemented by manual coding unless occupational data are complete and in the required format. Adjudication of coding discrepancies by an expert coder should yield the most accurate results, but this may not be practical for larger studies. We conclude that investigators should be concerned with coding reliability although this will be less of a concern when major categories or broader groups are of interest.

Acknowledgments

This research was supported by a National Institutes of Health grant [R01-HL081627]. The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute at the National Institutes of Health [contracts HHSN268201100005C, HSN268201100006C, HHSN268201100007C, HHSN268201100008C, HSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C]. The authors thank the staff and participants of the ARIC study for their important contributions.

References

Bushnell, D. An evaluation of computer-assisted occupation coding: Results of a field trial. Annual International Blaise Users Conference; Paris, France. 1997. p. 90-100.

Fleiss, JL.; Levin, B.; Paik, MC. Statistical methods for rates and proportions. New York: Wiley; 2003. p. 218

Kogevinas M. Commentary: Standardized coding of occupational data in epidemiological studies. Int J Epidemiol. 2003; 32:428–429. [PubMed: 12777431]

Mannetje A, Kromhout H. The use of occupation and industry classifications in general population studies. Int J Epidemiol. 2003; 32:419–428. [PubMed: 12777430]

NIOSH. Standardized occupation and industry coding: a software tool for automated coding of occupation and industry descriptions. 1.5. Morgantown, WV: National Institute for Occupational Safety and Health; 2001.

Ossiander EM, Milham S. A computer system for coding occupation. Am J Ind Med. 2006; 49:854–857. [PubMed: 16804909]

Rose KM, Perhac JS, Bang H, Heiss G. Historical records as a source of information for childhood socioeconomic status: Results for a pilot study of decedents. Ann Epidemiol. 2008; 18:357–363. [PubMed: 18395465]

The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. Am J Epidemiol. 1989; 129:687–702. [PubMed: 2646917]

U.S. Bureau of Census. 1990 Census of Population: Alphabetical Index of Industries and Occupations. Washington, DC: US Department of Commerce; 1992.

Table I

Distribution of Major Census Occupational Groups Assigned by Coder

	Manual Coder N=2,454		Manual Coder N=1,737[b]		Automated Coder[a] N=1,737	
	N	**(%)**	**N**	**(%)**	**N**	**(%)**
Non-manual Occupations						
Managerial and Professional Specialty occupations	219	(9)	182	(11)	185	(11)
Technical, Sales, and Administrative Support occupations	454	(19)	386	(22)	399	(23)
Manual Occupations						
Service occupations	123	(5)	97	(6)	96	(6)
Farming, Forestry, and Fishing occupations	124	(5)	102	(6)	113	(7)
Precision Production, Craft, and Repair occupations	540	(22)	425	(24)	443	(26)
Operators, Fabricators, and Laborers	880	(36)	508	(29)	469	(27)
Occupations Not Classifiable	19	(1)	16	(1)	20	(1)
Unknown Occupation	95	(4)	21	(1)	12	(1)

[a] Standardized Occupation and Industry Coding (SOIC) software

[b] Manual coding results for the subset coded by automated