

A computer system for coding occupation

Authors:

Eric M. Ossiander, MS (corresponding author)

Washington State Department of Health

P.O. Box 47812

Olympia WA 98504-7812, USA

voice: (360) 236-4252

fax: (360) 236-4245

email: eric.ossiander@doh.wa.gov

Samuel Milham, MD MPH

2318 Gravelly Beach Loop NW

Olympia WA 98502

Institution where work was performed:

Washington State Department of Health

Running head: A computer system for coding occupation

Abstract

Background Occupation information is widely used in epidemiologic studies and is collected on most death certificates and many birth certificates in the United States. Coding the massive amount of occupation information collected has been a challenge.

Methods A simple word-matching computer program to code occupation entries from vital records was developed. The accuracy of the program was evaluated by comparing its output to codes assigned by human coders.

Results In routine use in the Washington State Department of Health (DOH), the computer system codes 96–97% of the occupation entries on birth and death records. It assigned the correct code on 89% (95% confidence interval (87%, 91%)) of the records it coded.

Conclusions The occupation coding program is both efficient and accurate and can simplify the process of coding occupation entries from vital records. The system is adaptable and can be modified to use occupation classifications other than the one used by DOH.

Keywords: occupation, occupation classification, industry classification, vital records, computer occupation coding

Introduction

Occupation information is used in epidemiologic studies as an indicator of social class, to provide indirect adjustment for confounders, and as an indicator of occupational exposures (t Mannelje and Kromhout, 2003). In 1996, all states in the US collected occupation information on death certificates (Krieger et al., 1997) and the US standard death certificate includes it. In 1996, 25 states collected birth certificate occupation information for the mother and 24 collected occupation information for the father (Krieger et al., 1997), although the current US standard birth certificate does not include occupation information. Coding the massive amount of occupation information that is collected can be very difficult. Coding by hand is expensive, and the authors are not aware of any computer coding system, other than the one described here, which can code a high proportion of US vital records with acceptable accuracy.

Since 1992, all birth and death records filed at the Washington State Department of Health (DOH) have had the occupational information for the decedent (death certificate) and both parents (birth certificate) coded using the simple computer system described here. At DOH, the occupation and industry literals are keyed from the birth and death certificates and stored as part of the routine vital registration process. At present, the computer system codes approximately 96% of the birth and death records, with only 4% needing manual coding.

This paper describes how our computer coding system works, reports on a study of the accuracy of the codes obtained from the system, and provides some examples of how the coded data have been used in Washington State.

The occupation classification used at DOH is a modified version of the 1960 US Census classification, however, the computer system could be modified to use other classifications. The system provides a single code, rather than separate industry and occupation codes. This makes it easier to code a high proportion of the records, but makes the system less comparable to systems that produce two codes, such as the US Department of Labor's Standard Occupational Classification (SOC) and Standard Industry Classification (SIC) codes. The DOH classification has 3 digits and has 430 different occupation codes.

The computer programs and code dictionary used to implement the coding system described here are available on the Washington State Department of Health website (www3.doh.wa.gov/occmort).

Materials and Methods

The DOH computer coding system is a simple word matching system. The computer takes the individual words from the occupation and industry entries on the birth and death certificates, forms all possible permutations of the words, and matches the permutations against a dictionary. If the permutations that match do not all return the same occupation code, the

computer assigns a priority score based on the number of words that matched, and accepts the match with the highest score. If there is a tie among permutations with the highest priority score, then the computer evaluates those which are based on the occupation entry alone (not the occupation and industry entries). If there is still a tie, then the record is output for manual coding.

In Washington State, there are several prominent industries that are of particular interest, including aircraft manufacturing, aluminum manufacturing, and pulp and paper production. The system includes occupation code dictionaries specific to these industries, and all records matching these industries are coded separately. This is a feature of the program which can be easily modified or removed. The steps the program follows are described in Figure 1. Examples that illustrate how the system works are shown in Figures 2–5.

To evaluate the coding program a random sample of 800 certificates was drawn from files of birth and death certificates which had been previously coded by the computer program. The sample excluded records for children and records for which the computer program did not generate a code. Each of the authors coded half the sample, while blinded to the computer coding. A colleague listed the records for which the computer code and the code assigned by the authors did not agree. Then the samples were exchanged and these records were adjudicated. During the adjudication process, the authors were blinded as to which code was assigned by the computer. The

computer-assigned code was assumed to be correct if it agreed with the adjudicator, otherwise it was assumed to be incorrect.

The percentage of records coded correctly by the computer and by the authors was calculated, and the corresponding 95% confidence intervals. The percentage of records that the computer program is able to code during routine use at the Washington State Department of Health is also reported.

Results

Out of 800 records, the computer program coded 712 (89%, 95% confidence interval (87%, 91%)) correctly. The accuracy of the computer program nearly equaled that of the authors, who coded 714 (89%) of the records correctly. The coding accuracy of the computer program was similar between birth and death certificates (Table 1).

The occupations on the 88 records which the computer miscoded were well scattered across the list of occupations. Generally, the computer miscodes an occupation when some important words on the certificate are not in the computer's occupation dictionary, or when there are more words on the certificate that match to an incorrect code than words that match to a correct code. Here are some examples. The computer coded the entry "ARCHIECT/UNKNOWN" to the code for 'occupation unknown,' because it did not recognize the misspelled word for architect. The computer coded "LEGAL AGENT/SALES" to the code for 'sales agent' because that

matched more words than the single word ‘legal,’ and the phrase ‘legal agent’ was not in the dictionary. The computer coded “MICROFILM TECHNICIAN/ALUMINUM FABRICATION PLANT” to the code for aluminum worker because that matched more words than ‘microfilm technician.’ Although these miscodes can be corrected by adding entries to the spelling correction routine or the occupation dictionary, it is not possible to anticipate all the possible entries that would be needed to provide for completely accurate coding.

The program was used to code the 2003 birth and death certificates at the Washington State Department of Health. On the first pass, the program generated occupation codes for 97% of the death records, and for 96% of the fathers and 97% of the mothers on the birth records. These percentages include the codes that the computer generates for children or when the occupation entry on the certificate is left blank or stated as unknown.

Discussion

The computer system described here for coding the occupation entries on birth and death certificates is a simple system that works well. It codes 96–97% of records with an error rate of about 10%. DOH collects approximately 130,000 birth and death records each year. One author spends about 2 working days per year coding these records. Although this is not enough time to code all the records that the computer does not code,

it is enough time to update the system's code dictionary so that it continues to code 96–97% of the records each year. The program has been used to code all birth and death records at DOH since 1992, and a similar version on a mainframe computer was used for several years before that. Before this system was developed, coding occupation on just death certificates required 50% of a full-time staff person per year; DOH did not code occupation from birth certificates.

The coding program is written in SAS [SAS Institute, Cary, NC] and can easily be adapted to run on any computer that runs SAS version 6 or higher. Although the program has some features that are especially useful in Washington State, those features can be removed or modified. The system uses a code dictionary and includes routines for adding entries to the dictionary. Therefore, users who wish to use the same set of codes that we use, but in an environment in which some occupations are stated differently, can do this by updating the dictionary with new entries.

The system could also be modified to use a completely different set of occupation codes, while using the same programming logic, by creating a new coding dictionary.

The occupation codes provided by this system were used to conduct proportional mortality ratio (PMR) analyses of the relation between occupation and mortality in Washington State (Milham, 1997). The association between electrical occupations and leukemia was first noted in this data set (Milham, 1982). The latest analysis was conducted in 2001

and is available on the Washington State Department of Health web site (www3.doh.wa.gov/occmort). The web site displays PMRs for deaths to Washington residents occurring between 1950 and 1999 for men, and between 1974 and 1999 for women. For each occupation by cause-of-death grouping, PMRs were calculated for each calendar decade, and for the entire period; and for each 10-year age group, for the 20–64-year-old age group, and for all ages together. PMRs were computed separately for men and for women. The data display may be sorted by occupation, by cause-of-death, by the value of the PMR, by the value of the p-value associated with each PMR, and by combinations of these items.

The US National Institute for Occupational Safety and Health (NIOSH) has also developed a computer program for coding occupation and industry from death certificates and other records (NIOSH, 2005b). NIOSH reported that in an evaluation they conducted, their program agreed with an expert coder on 76% of the industry codes, 75% of the occupation codes, and on both codes 63% of the time (NIOSH, 2005b). The records on which the program and the coder disagreed were not adjudicated, and some of the disagreements may have been due to errors by the expert coder; however, it is likely that the accuracy of the NIOSH program is lower than that of the DOH program. Before adjudication, the agreement between the authors' codes and the computer's codes was only 78% in the comparison reported here, however, the authors are not expert coders. The NIOSH program is also used at DOH for coding occupation on death certificates. In routine

use at DOH, the program codes 75–80% of the occupation entries on death records; manual coding is required for the remaining 20–25% of records.

DOH does not use the NIOSH program for coding birth certificates because it does not have resources to do the manual coding that would be required.

The Office for National Statistics, in the UK, compared two different computer-assisted coding systems to expert human coders (Bushnell, 1997). One system was a knowledge-based system and the other used word-matching algorithms, as the DOH system also does. They found that using a word-matching system can be better than using a knowledge-based system, and that each system performed about as well as expert human coders (Bushnell, 1997). A review of several studies of occupation coding reliability conducted in the United Kingdom found that agreement rates between two human coders generally did not exceed 75% (Elias, 1997).

All states collect occupation information on death certificates and half of the states collect parental occupation information on birth certificates (Krieger et al., 1997). However, little of this information is coded into a form that is useful for researchers. In 1999, only 19 states used the NIOSH program to code death certificate occupation entries (NIOSH, 2005a).

A valuable feature of computer coding systems such as the DOH system is that they require keying the literal occupation entries into the computer. Once keyed, the literals can be stored and made accessible to researchers who require more detail than is available in the codes themselves. For

example, one of the present authors found an unusual sex ratio among births to fathers who work as aluminum plant carbon setters (Milham, 1993). Carbon setters form a small subset of workers who are coded to the aluminum worker rubric in the DOH coding system, and the unusual sex ratio of their children would not have been noted if the literal occupation entries had not been keyed.

Another advantage to keying literals is that the literals are available for coding to different coding systems and will be available for coding to new coding systems when those become available. Converting codes from one system to another can result in a loss of information (Kromhout and Vermeulen, 2001). Recoding from stored literals will avoid this problem. Routine keying of occupation and industry literals from all vital records is recommended.

References

- 't Mannetje A, Kromhout H, 2003. The use of occupation and industry classifications in general population studies. *Int J Epidemiol* 32:419–28.
- Bushnell D, 1997. An evaluation of computer-assisted occupation coding: Results of a field trial. Paris, France, pp. 90–100. Annual International Blaise Users Conference, 1997.
- Elias P, 1997. Occupational classification: concepts, methods, reliability, validity and cross-national comparability. Institute for Employment Research, University of Warwick.
<http://www.lisproject.org/publications/leswps/leswp5.pdf>; accessed April 12, 2005.
- Krieger N, Chen JT, Ebel G, 1997. Can we monitor socioeconomic inequalities in health? A survey of U.S. health departments' data collection and reporting practices. *Public Health Rep* 112:481–91.
- Kromhout H, Vermeulen R, 2001. Application of job-exposure matrices in studies of the general population: some clues to their performance. *Eur Respir Rev* 11:80–90.
- Milham S, 1982. Mortality from leukemia in workers exposed to electrical and magnetic fields. *N Engl J Med* 307:249.

Milham S, 1993. Unusual sex ratio of births to carbon setter fathers. *Am J Ind Med* 23:829–31.

Milham S, 1997. Occupational mortality in Washington State, 1950-1989. DHHS Publication No.96-133, NIOSH.

NIOSH, 2005a. Occupational respiratory disease surveillance. URL <http://webappa.cdc.gov/ords/norms-states.html>. National Institute for Occupational Safety and Health. Accessed April 4, 2005.

NIOSH, 2005b. Standardized occupation & industry coding. URL <http://www.cdc.gov/niosh/SOIC/>. National Institute for Occupational Safety and Health. Accessed March 29, 2005.

1. Read the data. Need these items: Unique identifier, occupation literal, industry literal (optional), and age.
2. Extract and set aside records for children and records where both the occupation and industry literals are stated as unknown or left blank.
3. Remove punctuation from the occupation and industry literals and break them into separate words.
4. Standardize the word spellings.
5. Form all ordered permutations of the words, and output those that match an entry in the occupation dictionary. Along with each permutation, output a priority score that consists of the number of words in the permutation.
6. Check the permutations against the housewife table and the special industry tables. Output the records that matched into a special industry file.
7. Check the remaining records against the general occupation file.
8. For records not in the special industry file, repeat the process using only the occupation literals.
9. If a record has matched to more than one code from the occupation code tables, then use the priority score to determine which code to assign to that record. If there is a tie between a permutation derived from the occupation and industry literals and a permutation derived from the occupation literal alone, then assign the code derived from the occupation literal alone. If there is still a tie, do not assign a code.

Figure 1: Procedure the program uses to code occupation.

Permutations which matched dictionary	Priority score	From occupation only?	Occupation code
FUEL OIL TRUCK DRIVER	4		716
OIL TRUCK DRIVER	3		716
TRUCK DRIVER FUEL	3		716
FUEL OIL	2		295
FUEL TRUCK	2		716
TRUCK DRIVER	2		715
TRUCK DRIVER	2	Y	715
DRIVER	1		715
FUEL	1		291
DRIVER	1	Y	715

Figure 2: The occupation entry was “truck driver,” and the industry entry was “fuel oil.” Eight permutations of these four words matched an entry in the occupation dictionary. Four permutations of the two words in the occupation entry also matched to the dictionary. Only one match had the highest priority score of four, therefore the code from that match was accepted.

Permutations which matched dictionary	Priority score	From occupation only?	Occupation code
CONSTRUCTION	1		982
MACHINIST	1		465
MACHINIST	1	Y	465

Figure 3: The occupation entry was “machinist,” and the industry entry was “construction.” Two permutations of these two words matched an entry in the occupation dictionary. The one word in the occupation entry also matched to the dictionary. Each match had the same priority score of one. Therefore, the match which used the occupation entry alone was used to assign the code.

Permutations which matched dictionary	Priority score	From occupation only?	Occupation code
CONSTRUCTION FRAMER	2		411
CONSTRUCTION FRAMER	2		411
CONSTRUCTION FRAMER	2	Y	411
CONSTRUCTION	1		982
CONSTRUCTION	1		982
FRAMER	1		411
CONSTRUCTION	1	Y	982
FRAMER	1	Y	411

Figure 4: The occupation entry was “construction framer,” and the industry entry was “construction.” Five permutations of these three words matched an entry in the occupation dictionary. Three permutations of the two words in the occupation entry also matched to the dictionary. Three matches had the highest priority score, but all three matched to the same occupation code, so that code was accepted.

Permutations which matched dictionary	Priority score	From occupation only?	Occupation code
HEATING CONTRACTOR	2		470
PLUMBING CONTRACTOR	2		510
CONTRACTOR	1		406
HEATING	1		470
PLUMBING	1		510
CONTRACTOR	1	Y	406

Figure 5: The occupation entry was “contractor,” and the industry entry was “plumbing & heating.” Five permutations of these three words matched an entry in the occupation dictionary. The one word in the occupation entry also matched to the dictionary. Two matches had the highest priority score, but were matched to different occupation codes. The match with the occupation entry alone had a lower score, so it was not considered. No code could be assigned.

Table 1: Coding accuracy of the computer occupation coding system.

Type of record	N	Coded		
		correctly	Percent	95% CI
Birth	400	364	91	(88, 94)
Death	400	348	87	(84, 90)
Total	800	712	89	(87, 91)