ARIC Manuscript Proposal # 1546

PC Reviewed: 8/11/09	**Status: A**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title:

Performance of Manual and Automated Occupation Coding Systems with Historical Occupational Data

b. Abbreviated Title (Length 26 characters): Occupational Coding

2. Writing Group:

Writing group members: Mehul D. Patel, Kathryn M. Rose, Heejung Bang, Jay S. Kaufman, and others welcome

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. __MDP___ **[please confirm with your initials electronically or in writing]**

First author: **Mehul D. Patel, MSPH**
Address: 137 E. Franklin Street
Bank of America Center, Suite 306
Chapel Hill, NC 27514

Phone: (919) 260-1651 Fax:
E-mail: mdpatel@email.unc.edu

ARIC author to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).

Name: **Kathryn M. Rose**
Address: 137 E. Franklin Street
Bank of America Center, Suite 306
Chapel Hill, NC 27514

Phone: (919) 966-4596 Fax: (919) 966-9800
E-mail: Kathryn_rose@unc.edu

3. Timeline:

Analysis to begin in the Fall of 2009

4. Rationale:

Background

In the U.S., birth certificates have been required by state law since the 1930's, and although not standardized across states, these early records typically contain parents' occupation[1]. The U.S. Census Bureau has been collecting occupation on standard individual records since 1840[2]. These records are declassified after 72 years by law (Title 44, U.S. Code); thus, records from the 1930 U.S. Census have been available to the public since 2002. Two recent epidemiological studies used historical records as a source of childhood socioeconomic information. Moceri et al. found a census record and/or birth certificate for 86% of elderly subjects in a study of early life environment and Alzheimer's disease[3]. Rose et al. searched for records on decedents born in North Carolina (NC) from 1921-1935 and located a NC birth certificate and/or 1930 census record for 85%[4]. For decedents with father's occupation from both birth certificate and census record, occupation was classified into six census-based categories, and the observed agreement between the two sources was 89%.

Epidemiological studies that collect information on occupations typically categorize these data to be used as indicators of occupational exposures or SES. The U.S. Census Bureau uses an index of 501 occupations to classify job titles and to group them into seven major occupational categories. This system offers a standardized approach to coding occupations and can improve the comparability of occupation-based measures between epidemiological studies[5].

Rationale

Typically, trained coders are employed to manually assign standardized census codes to occupations. This method can be time- and labor-intensive, and there is the potential for variability between and within coders to reduce the reliability of occupational information. However, these problems are rarely evaluated. Also, in recent years, skilled, rigorously trained coders are becoming increasingly difficult to find.

Automated computer programs and computer-assisted coding are viable alternatives[5-8]. The National Institute for Occupational Safety and Health (NIOSH) has developed the Standardized Occupation and Industry Coding (SOIC) system - a free of charge, standalone Windows-based software designed to read occupation and industry narratives and assign 3-digit numerical occupation and industry codes based on the 1990 U.S. Census Bureau Index of Industries and Occupations[9]. Computer programs, such as the SOIC, are designed to consistently code occupational data to standard classification systems and can make the coding process more efficient for researchers. However, the performance of such programs has not been sufficiently studied. NIOSH compared results from the SOIC with a manual coder and found 75% agreement between occupation codes, but the accuracy of computer codes was not further evaluated. In some preliminary work done in our group, a major source of the disagreement between manual and computer assigned codes was a higher proportion of not being able to assign a code by the computer program[10].

We aim to compare coding performances of an automated computer coding program, a recently-trained ("novice") manual coder, and an experienced, NIOSH-certified manual coder and to assess the agreement of assigned occupational codes between these three coding methods. We will also assess the consistency of coding within the two manual coders by having them recode a sample of original occupation data. Additionally, the validity of the codes assigned by the computer and the novice manual coder will by estimated by comparing them to codes assigned by the "gold standard" experienced coder. Based on the results of this reliability and validity study, we plan to make a recommendation for the use of automated and manual occupation coders in epidemiologic studies.

5. Main Hypothesis/Study Questions:

We will consider the following three occupational coding methods:

- Automated computer coder
- Novice manual coder
- Experienced manual coder

For each coder, we will quantify:

1. Percent of records assigned a 3-digit occupation code based on the 1990 U.S. Census Bureau Index of Industries and Occupations
2. Differences in records assigned a code to those not assigned a code by:
    - Source (birth certificate or census record)
    - Study community (i.e. Forsyth County, NC; Jackson, MS; northwest suburbs of Minneapolis, MN; Washington County, MD)
    - Sex
    - Black or White race

We will evaluate the agreement of the following between the three coding methods:

1. 3-digit occupation code
2. Major category
    - Census occupational categories:
        i. Managerial and Professional Specialty Occupations
        ii. Technical, Sales and Administrative Support Occupations
        iii. Service Occupations
        iv. Farming, Forestry and Fishing Occupations
        v. Precision Production, Craft, and Repair Occupations
        vi. Operators, Fabricators, and Laborers
        vii. Military Occupations
    - Not Applicable (i.e. homemaker, student, unemployed, retired, disabled)
    - Blank/Unknown (entries left blank or stated as unknown)

For the two manual coders, we will evaluate and compare the consistency of assigning occupation codes and categories between two separate submissions.

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodologic limitations or challenges if present).

The ARIC ancillary study Life Course SES, Social Context and Cardiovascular Disease (LCSES) collected parental occupational information from participants at a follow-up visit (2001-2002). Information on the parent's occupational category was obtained with a telephone questionnaire. Participants were asked to select which of eight census-based major occupation categories describes the type of work their father (or male caretaker) and mother did when they (the participants) were children. An ancillary ARIC study entitled "Using Historical Records to Reconstruct Early life SES Exposures of Decedents" (Rose, PI) is collecting father's occupation and industry titles from the birth certificates and parents' census records of 3,444 ARIC decedents for the purposes of (1) obtaining early life SES data on ARIC decedents who died prior to participating in the LCSES study (2) assessing the extent of recall error and survivorship bias in LCSES results.

For this analysis, we will use father's occupational and industry data abstracted from all birth certificates and census records located for ARIC decedents. Occupation and industry titles of the participants' fathers will be coded by an automated computer system: the Standardized Occupation and Industry Coding (SOIC) system[9]. Records will also be independently coded to the 1990 U.S. Census Bureau Index of Industries and Occupations by a novice coder (the first author) and an experienced coder (hired). Each manual coder will recode a 10% random sample of records to determine the consistency within each coder. The time taken to complete the coding will be recorded by the manual coders.

We will compare the performance between coders using the percent of records assigned a code. We will also assess whether records with no code assigned are associated with specific factors including the type of record and the participant's community, sex, and race. Coding results will be compared with the assigned 3-digit occupation code and with results grouped into the seven major occupational categories. Percent agreement and the kappa statistic, with 95% confidence intervals, will be used to measure inter-coder reliability[11-15]. The same reliability measures will be used to evaluate consistency within each manual coder. In general, percent agreement at 90% or above and kappa statistics at 0.80 or higher will be considered very good and acceptable. Since the experienced coder has been trained and certified and has over 25 years of professional industry and occupation coding experience, she will be considered the gold standard, and we will assess the accuracy of the computer and novice coders by comparing their assigned occupation codes and categories to the experienced coder's.

7.a. Will the data be used for non-CVD analysis in this manuscript? __x__ Yes ____ No

b. If Yes, is the author aware that the file ICTDER03 must be used to exclude persons with a value RES_OTH = "CVD Research" for non-DNA analysis, and for DNA analysis RES_DNA = "CVD Research" would be used? ____ Yes __x__ No
(This file ICTDER03 has been distributed to ARIC PIs, and contains the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript? ____ Yes __x__ No

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the file ICTDER03 must be used to exclude those with value RES_DNA = "No use/storage DNA"?
____ Yes ____ No

9.The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status. ARIC Investigators have access to the publications lists under the Study Members Area of the web site at: http://www.cscc.unc.edu/ARIC/search.php

___x___ Yes _____ No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?
MS970

11. a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? __x__ Yes ____ No

11.b. If yes, is the proposal
_x__ A. primarily the result of an ancillary study (list number* _2003.07__)
___ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* _____ _____ _____)

*ancillary studies are listed by number at http://www.cscc.unc.edu/aric/forms/

12. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

References

1. Shapiro S. Development of birth registration and birth statistics in the United States. Population Studies: A Quarterly Journal of Demography 1950:86-111.
2. 200 years of U.S. census taking: population and housing questions, 1790-1990. In: Census USBot, ed. Vol. iv. Washington, D.C., 1989;109.
3. Moceri VM, Kukull WA, Emanual I, Belle Gv, Starr JR, Schellenberg GD, McCormick WC, Bowen JD, Teri L, Larson EB. Using Census Data and Birth Certificates to Reconstruct the Early-Life Socioeconomic Environment and the Relation to the Development of Alzheimer's Disease. Epidemiology 2001;12(4):383-389.
4. Rose KM, Perhac JS, Bang H, Heiss G. Historical Records as a Source of Information for Childhood Socioeconomic Status: Results from a Pilot Study of Decedents. Annals of Epidemiology 2008;18(5):357-363.
5. 't Mannetje A, Kromhout H. The use of occupation and industry classifications in general population studies. International Journal of Epidemiology 2003;32(3):419-428.
6. Bushnell D. An Evaluation of Computer-Assisted Occupation Coding: Results of a Field Trial. Annual International Blaise Users Conference. Paris, France, 1997;90-100.
7. Kogevinas M. Commentary: Standardized coding of occupational data in epidemiological studies. Int. J. Epidemiol. 2003;32(3):428-429.
8. Ossiander EM, Milham, Samuel. A computer system for coding occupation. American Journal of Industrial Medicine 2006;49(10):854-857.
9. National Institute for Occupational Safety and Health DoSR. Standardized Occupation and Industry Coding. 1.5 ed. Morgantown, WV, 2001;a software tool for automated coding of occupation and industry descriptions.
10. Patel MD, Rose KM, Owens CR, Bang H, Kaufman JS. Evaluation of Automated Coding to Classify Occupations on Historical Records. [Abstract] American Journal of Epidemiology 2008;167(11):S127.
11. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 1960;20(1):37-46.
12. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin 1971;76(5):378-382.
13. Koch GG, Landis JR, Freeman JL, Freeman DH, Jr., Lehnen RG. A General Methodology for the Analysis of Experiments with Repeated Measurement of Categorical Data. Biometrics 1977;33(1):133-158.
14. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics 1977;33(1):159-174.
15. Landis JR, Koch GG. An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. Biometrics 1977;33(2):363-374.