UK Biobank

Methods used to derive job codes from free-text job descriptions

Version 1.0

http://www.ukbiobank.ac.uk/
June 2013

This document outlines the procedure for coding participants’ job descriptions, which were entered as free text during the verbal interview stage of the assessment centre visit.

Contents

1 Introduction

1.1 Participants were asked about their current employment status in the touchscreen questionnaire section (2nd station, table 1) of the assessment centre visit. Those who had indicated they were employed or self-employed were subsequently asked during the verbal interview stage (3rd station) to describe their current job in more detail.

1.2 The job descriptions provided were coded at the time using Standard Occupational Classification (SOC) 2000 codes, or entered by the interviewer as free text, if a suitable code could not be assigned. Further details are provided elsewhere: http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=100235.

1.3 The information on employment obtained during the touchscreen and verbal interview stage is contained in data fields 6142 *Current employment status* and 132 *Job code,* respectively. However, the *Job code* data-field only includes those jobs that were able to be coded by the interviewer at the time; it does not include detailed information on participants who had their job descriptions entered as free text.

1.4 This document outlines the process of coding the free-text job descriptions. This work was performed by Dr Sara De Matteis, Dr Lesley Rushton and Professor Deborah Jarvis, Department of Epidemiology and Biostatistics School of Public Health, Imperial College London.

Table 1: Sequence of assessment visit

	Visit station	**Assessments undertaken**
1	Reception	• Welcome & registration • Generating a USB key for Participants
2	Touch screen Section	• Consent • Touch screen questionnaire • Hearing Test • Cognitive function tests
3	Interview & blood pressure	• Interviewer questionnaire • Blood pressure measurement • Measurement of arterial stiffness
4	Eye measurements	• Visual acuity • Refractometry • Intraocular pressure • Optical Coherence Tomography
5	Physical measurements	• Height (Standing and Sitting) • Hip & Waist measurement • Weight and Bio-impedance measurement • Hand-grip strength • Ultrasound Bone Densitometry • Spirometry (Lung function test)
6	Cardio-respiratory fitness test	• Exercise/fitness ECG test
7	Sample collection & exit	• Blood samples collected • Urine sample collected • Saliva sample collected
8	Web-based diet questionnaire	• Dietary assessment

2 Methods

2.1 Data Cleaning

2.1.1 Some free-text descriptions contained text that was not applicable for coding and all job descriptions were first screened for the following key words: *retired* or *retied*; *voluntary* or *unpaid*; *house* or *home* or *family*; *employ* or *unemployed* or *unemployment* or *none* or *not employed*; *sick* or *ill* or *disease* or *handicap* or *disable*; *not respondent* or *unwilling* or *refuse* or *answer*; *student* or *Phd* or *MSc*.

2.1.2 Job descriptions were not coded for any participant who was identified as not being in current employment, as defined above (although their original free-text response is retained for our records).

2.2 CASCOT coding

2.2.1 A Computer Assisted Structure COding Tool (CASCOT), developed by the University of Warwick[1], was used to translate the free text job descriptions into Standard Occupational Classification (SOC) 2000 codes.

2.2.2 The CASCOT programme generates a confidence score (1-100%) for each job description, which represents the Bayesian probability that the computer-assigned SOC code is that which would be assigned manually by experts in job coding.

2.2.3 The CASCOT programme was performed on the free text job descriptions, both with and without correction for mistypes and abbreviations.

2.2.4 The results of the CASCOT programme, based on the free text job descriptions with correction, were retained for incorporation into to the UK Biobank resource.

3 Data obtained and presented in UK Biobank

3.1 Of the 18,322 free-text entries, there were 12,826 unique job descriptions, which underwent analysis.

3.2 1,019 (7.9%) unique free-text job descriptions were not applicable for coding because they had self-reported that they were not in current employment (i.e. retired, unable to work, performing voluntary work, etc.). This figure rose to 1,051 (8.2%) after correction of the free-texts for mistypes and abbreviations.

3.3 99.8% (n=11,751) of the remaining free-text job descriptions were successfully translated into a Standard Occupational Classification (SOC) 2000 code, using the CASCOT programme.

3.4 The CASCOT programme performed slightly better after the job descriptions were corrected for mistypes and abbreviations, with the proportion of confidence scores ≥50% rising from 62.7% to 64.1%.

3.5 Given the higher performance of the CASCOT programme, the data available in the UK Biobank resource is based on the corrected job descriptions.

3.6 Of the 18,322 participants who had free-text job description entries, 3,366 (18.4%) participants were identified as not being in current employment. The *Current employment status* data-field (6142) has therefore been updated for these participants to reflect these changes (data field 20119 *Current employment status – corrected*), as shown in table 2.

Table 2: Coding for data field 20119: *Current employment status – corrected*

Code / format	**Description**
1	In paid employment or self-employed
2	Retired
3	Looking after home and/or family
4	Unable to work because of sickness or disability
5	Unemployed
6	Doing unpaid or voluntary work
7	Full or part-time student
-7	None of the above
-3	Prefer not to answer

3.7 14,568 (79.5%) participants in current employment have now been assigned a Standard Occupational Classification (SOC) 2000 code. These are contained in data field 20024 *Job code – deduced* (sister code for 132), along with the corresponding confidence score 20121 *Confidence score.*

3.8 The original free-text job descriptions used in this exercise (data-field 20120) are available on request.

4 References

1) The Warwick Institute for Employment Research website. (http://www2.warwick.ac.uk/fac/soc/ier/software/cascot/)