Scotland's Census 2011

Data quality issues for the long-term health conditions question

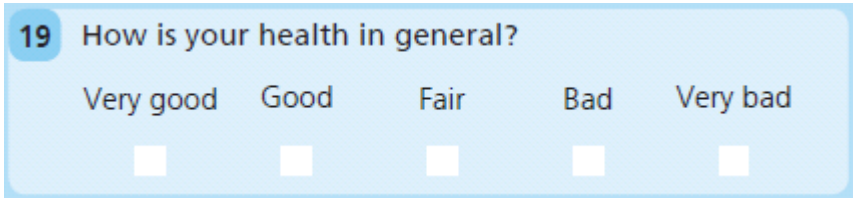
13 November 2017

A National Statistics publication for Scotland

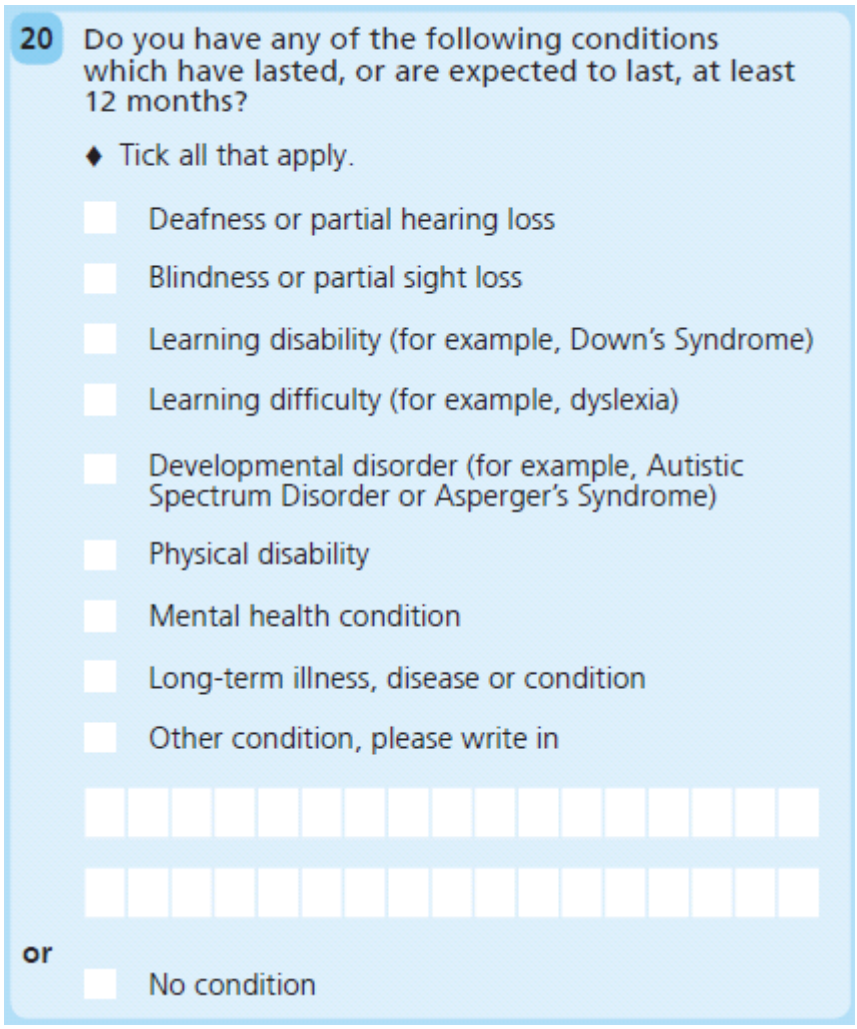
National Statistics are produced to high professional standards set out in the National Statistics Code of Practice. They undergo regular quality assurance reviews to ensure that they meet customer needs.

1. Introduction

1.1 This report provides background details on some of the data quality issues around the Long-Term Health Conditions questions in Scotland's Census 2011. It is mostly concerned with the results of question 20 ("Do you have any of the following conditions which have lasted, or are expected to last, at least 12 months?"), but some information about question 19 ("How is your health in general?"), and question 21 ("Are your day-to-day activities limited because of a health problem or disability which has lasted, or is expected to last, at least 12 months?") is also included as these question have been used to adjust the results for the Long-Term Health Condition variables.


19 How is your health in general?

Very good ☐ Good ☐ Fair ☐ Bad ☐ Very bad ☐


20 Do you have any of the following conditions which have lasted, or are expected to last, at least 12 months?

♦ Tick all that apply.

☐ Deafness or partial hearing loss

☐ Blindness or partial sight loss

☐ Learning disability (for example, Down's Syndrome)

☐ Learning difficulty (for example, dyslexia)

☐ Developmental disorder (for example, Autistic Spectrum Disorder or Asperger's Syndrome)

☐ Physical disability

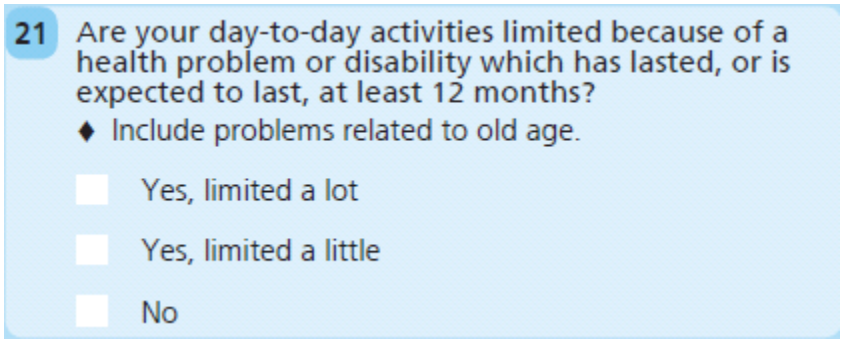
☐ Mental health condition

☐ Long-term illness, disease or condition

☐ Other condition, please write in

or

☐ No condition



1.2 The derived results of question 20 are stored in the census database as the variable LTCOND1. This variable is 10 digits long, with each digit representing a tick box in question 20 in the same order as in the question (ie. Tick box 1 (Deafness or partial hearing loss) is digit 1 in the variable). The only possible values for the digits are 0 or 1 for the final variable, with 0 meaning that the respondent does not have the condition, and 1 meaning that the respondent does have the condition. For example, the value “0010100000” would mean that the respondent reported that they had a Learning Disability and a Developmental Disorder”. If the respondent reported that they had “No condition”, then this was recorded as “0000000001”.

1.3 The census database also includes a variable LTCOND2, which records the number of categories of long-term condition reported for a person. (A further variable - LTCOND3 – was held on earlier iterations of the census database to record write-in responses to the free text box in question 20. However, this variable was removed from later versions of the database).

1.4 During processing of the census data, one filter rule (Filter Rule 13) and one Data File Amendment (DFA 69) were applied which altered the initial values of LTCOND1. Further details on these are given in section 2.

1.5 When a value for a variable in the census database is changed, eg when a value is imputed for a missing response, then an imputation flag is set for that variable. Imputation flags for LTCOND1 are stored as the variable LTCOND1_imp_output. This imputation flag isn’t suitable for informing analysis as any change to any of the 10 digits in LTCOND1 will be recorded for the entire variable. It isn’t possible to tell which conditions were altered using this imputation flag.

2. Changes to LTCOND1

2.1 **Filter rule 13** altered the results of LTCOND1 for some respondents who missed the question. A consequence of the structure of question 20 meant that it was easy for a respondent who did not have a long-term health condition to miss the final “No Condition” text box. Filter rule 13 was designed to assign “No Condition” to such people if their answers to question 19 and question 21 made it likely that they are generally health. If a respondent who reported that their health was “Very Good” or “Good” (1 or 2) and that they did not have a limiting disability, then LTCOND1 was set to '0000000001', i.e. that they had no long-term health condition.

2.2 **Data File Amendment 69** was applied to the census database at a later stage. This was intended to adjust LTCOND1 to account for any text that had been entered in question 20 (stored as the variable LTCOND3). All unique entries in LTCOND3 were hand coded to record which of the tick boxes in question 20 they corresponded to best. LTCOND1 was then amended so that, if a person had written in a valid condition, the digit that corresponded to this in LTCOND1 was altered to '1'. For instance, if a respondent had original values for LTCOND1 of "0010000000" and for LTCOND3 of "I am deaf", then LTCOND1 would be altered to "1010000000".

2.3 **New LTCOND variables** were required for analytical purposes as it was necessary to be able to tell the difference between people who were imputed as having particular conditions, and people who had stated on their form that they had a condition or conditions. It was not possible to make this distinction using just the imputation flag variable LTCOND1_imp_output.

2.4 It was decided that the best way to store this data in the census database was to split the variable LTCOND1 into eight separate variables corresponding to each of the eight long-term health condition categories in the census questionnaire. (It was judged that for the purposes of analysis there was no valid distinction between responses for the "Other Condition" and "Long-term illness, disease or condition" tick boxes, and so the two were merged into a single "other condition" category.) variable.) "No Condition" was not directly relevant to these new variables, so no new variable was built for it.

2.5 Each of these new variables had a corresponding imputation flag variable created for it. These were used to record where the relevant information for a person was altered during processing. This was useful in identifying which records related to people who had a specific category of long-term health condition. For analysis purposes the aim was that records altered by DFA 69 to show a person as having have a specific category of long-condition health condition should be treated the same as if they had originally ticked the correct box in question 20.

3. Quantification of quality issues in LTCOND variables

3.1 It was assumed that most people who missed question 20 generally did so because they had no long-term health condition but didn't spot the "No Condition" check box in the question. While there is not believed to be any user-tested evidence on this, this seems to be the most plausible explanation.

3.2 There were 5,071,563 records in the tick and text files created by the initial data capture process which could be matched successfully to a record in the census database. Of these, there were 702,221 people (14.7% of the total) who had missed out question 20. Of these people, there were 569,654 people who also met the condition for filter rule 13, of which 97.4% had been recoded as LTCOND1 = "0000000001" (No condition). That this was not 100% is probably because of record swapping after filter rule 13 was run.

3.3 For those people who did correctly tick/enter text into question 20, and met the same condition as Filter Rule 13, only 87.2% (3,288,011) reported that they had "No Condition".

3.4 If the rate of people without long-term health conditions meeting the rules for filter rule 13 is accurately represented by the people referred to in section 3.3, then the number of people without a long-term health condition in Scotland would be inflated by around 90,000 people. If the rate is significantly less then we have no way of knowing what number of people have been misreported.

Data File Amendment (DFA) 69

3.5 There were 387,264 respondents where LTCOND1 was modified by DFA 69 based on the write-in responses recorded by variable LTCOND3.

3.6 Conditions could only be coded in DFA 69 if LTCOND3 could be interpreted correctly. Text written in question 20 was recorded by computer vision, so there are many examples of letters being mis-recorded by LTCOND3 from the census questionnaire. Some of the images for census questionnaires where values of LTCOND3 seems to be poorly recorded by the software have been reviewed. This revealed that some of these questionnaires had extremely shaky and scratchy handwriting, which would be much more difficult for the software used to interpret correctly, and as a result be less likely to be correctly coded during DFA 69.

3.7 A possible consequence of this is that there will be long-term health conditions that adversely affect handwriting, and these will be less likely to be recorded and coded correctly from the DFA. However, there is currently no quantification of this potential issue.