Australian Longitudinal Study
on Women's Health

Notes compiled for collaborators

Contents

Sampling scheme for the 1921-26, 1946-51, 1973-78, and 1989-95 cohorts

Selection of the sample

1973-78, 1946-51 and 1921-26 cohorts

The study sample was selected by Medicare Australia (previously known as the Health Insurance Commission) from three zones - urban, rural and remote defined according to the Australian Standard Geographical Classification RRMA scheme where urban includes Capital City and Other Metropolitan Centres; rural, Large Rural Centres, Small Rural Centres and Other Rural Areas; and remote, Remote Centres and Other Remote Areas.

The age groups sampled from the Medicare database in April 1996 were 18-22 years, 45-49 years and 70-74 years. By the time the invitations to participate were mailed later in 1996, some women at the upper limit of the age groups had had their birthday and were a year older. Hence, some women recruited were 23, 50 and 75 years old and so the cohort age ranges in the study are: 18-23, 45-50, and 70-75 years (although there are relatively fewer women in the oldest year of each cohort). The cohorts are now referred to by their years of birth but some study material may refer to them as 'Young', 'Mid-aged' and 'Older' and datasets use 'y', 'm', and 'o' (further information below).

Sampling from the population was random within each age group, except that women from rural and remote areas were selected in twice the proportions of the Australian population living in these areas. Women from capital cities and other metropolitan areas made up the balance of the samples.

There were also a small number of women invited to participate whose age was outside the cohort birth years (by a year or two), possibly due to errors in date of birth in the Medicare database. However, the survey data for these women have been retained. We recommend that when using the data, these women are either excluded or their age set to the nearest valid age.

1989-95 cohort

Please note that some variables in Surveys 1 and 2 of the 1989-95 cohort were renamed for consistency in April 2016.

See: Renaming of Variables in the Surveys 1 and 2 for the 1989-95 cohort

Recruitment for the 1989-95 cohort was different from the other cohorts. A variety of recruitment strategies were used (see the Report I, section 3.) A brief summary is given here.

For inclusion in the 1989-95 cohort, respondents needed to:

- meet the eligibility criteria of being female, aged 18 to 23 and having a Medicare number;
- answer at least some survey questions; and
- meet the requirements for data linkage.

A total of 17,567 women met the above inclusion criteria. To establish a pilot study group for the cohort, the first 498 young women that met the above criteria were removed from the main cohort. As a result, the pilot study group included all women recruited in October 2012 who were verified by the Department of Human Services. Of the remaining sample, 17,069 participants were verified by the Department of Human Services.

Some participants in this cohort were later found to be ineligible due to their birth years being out of range and they have been removed from the cohort. In April 2018, there were 17010 participants in the cohort.

Calculation of the sample weights

1973-78, 1946-51 and 1921-26 cohorts

The women were selected based on their postcodes recorded by Medicare. The variable in the datasets called 'inarea' reflects the area from which the women were sampled (urban, rural, remote). However by the time the survey was mailed, some women, particularly in the younger age group, had moved. The variable 'y1area' reflects their actual area of residence when completing the survey. The number of respondents who lived in urban, rural and remote areas at the time of completing the first survey in 1996 (wave 1 area) was used to create the sample weights for each age group for each area (urban, rural, remote), by comparing these numbers of respondents to 1996 Census figures. The sample weights appear in the datasets and are labelled y1wtarea, m1wtarea and o1wtarea.

1989-95 cohort

Sample weights were calculated for the 1989-95 cohort based on the women’s ages and areas of residence (urban, rural and remote). The 2011 Census was used as the best available measure of Australia’s population of women aged 18 to 23.

Weights for women in the sample of age *x* (at baseline) residing in geographical region *z:*

$(x, z) = [P(x, z)/P] \div [N(x, z)/N]$

Where N is the total number of women in the sample and N(x, z) is the number of women aged x years residing in geographical region z in the sample. Similarly, P is the total number of women aged 18 to 23 in the Australian population, and P(x, z) is the number of women in the Australian population aged x years residing in geographical region z.

Representativeness and attrition

These papers explain representativeness and attrition:

- Lee C, Dobson AJ, Brown WJ, Bryson L, Byles J, Warner-Smith P, Young AF. (2005) Cohort Profile: The Australian Longitudinal Study on Women's Health. *International Journal of Epidemiology*; 34: 987-991.
- Young AF, Powers JR, Bell SL. Attrition in longitudinal studies: who do you lose? *Australian and New Zealand Journal of Public Health*. 2006 Aug;
- Brilleman SL, Pachana NA, Dobson AJ. The impact of attrition on the representativeness of cohort studies of older people. *BMC Medical Research Methodology*. 2010 Aug;10.
- Powers J, Loxton D. The Impact of Attrition in an 11-Year Prospective Longitudinal Study of Younger Women. *Annals of Epidemiology* 2010; 20(4):318-21.)

For representativeness for the 1989-95 cohort see:

- Health and wellbeing of women aged 18 to 23 in 2013 and 1996: Findings from the Australian Longitudinal Study on Women’s Health. Mishra G, Loxton D, Anderson A, Hockey R, Powers J, Brown W, Dobson A, Duffy L, Graves A, Harris M, Harris S, Lucke J, McLaughlin D, Mooney R, Pachana N, Pease S, Tavener M, Thomson C, Tooth L, Townsend N, Tuckerman R and Byles J. Report prepared for the Australian Government Department of Health, June 2014. (Section 4).

Longitudinal analysis

When doing longitudinal analyses with the cohorts beginning in 1996, remember to weight for area of residence at Survey 1 (y1wtarea, m1wtarea, o1wtarea) in all crosstabs, frequencies and analyses to adjust for the initial deliberate oversampling in rural and remote areas. This weighting may not be required in models that include a geographic area of residence variable. For information on geographic area of residence, see below in Notes about specific variables.

Missing data

Some participants completed a short survey instead of the full survey, accounting for some missing data. The type of survey completed is identified with variables such as y2survey for Survey 2 of the 1973-78 cohort. Survey 2 of the 1946-51 cohort Q70 on income is missing the first category ($1-$119). There are large amounts of missing data in some income questions. Surveys 2, 3 and 4 of the 1946-51 cohort are missing the question about being admitted to hospital. Survey 2 of the 1973-78 cohort is missing the question about ability to manage on income. Survey 2 of the 1946-51 cohort Q67 is unreliable as the instruction was incorrectly stated as "mark one only" rather than "mark all that apply.” Many participants realised that this was an error and answered the question, as it should have been. Others may not have done so.

The first survey of the 1989-95 cohort has 167 records whose data are almost all missing. These records are identified by the *allmissing* variable. This variable has the value 1 for those records that are almost all missing, zero otherwise. These records represent eligible respondents who did complete the first survey but we unfortunately lost their data. They are kept in the dataset so that the first wave’s dataset contains the whole sample.

Notes about data files

The quantitative survey data are available as SAS, STATA and SPSS data files, or as tab delimited text files. The dataset files include almost all survey items as well as all derived and calculated variables.

Naming conventions for datasets

The analysis datasets without formats and labels attached are named WHA*survey*cohortB

Where *survey* is the survey wave number

Where ‘cohort’ is the three-letter cohort abbreviation:

yng (1973-78 cohort), mid (1946-51 cohort) , old (1921-26 cohort), nyc (1989-95 cohort)

B = level B data (identifying information removed). For example, wha1yngB.txt is the text dataset for Survey 1 of the 1973-78 cohort.

The analysis datasets with formats and labels attached are named W*survey*cohortBF

Where *survey* is the survey wave number

Where ‘cohort’ is the one letter cohort abbreviation:

y (1973-78 cohort), m (1946-51 cohort) , o (1921-26 cohort), z (1989-95 cohort)

B = level B data (identifying information removed) and F refers to formats and labels attached. For example, w2mBf.sas is the SAS dataset with formats for Survey 2 of the 1946-51 cohort.

Naming conventions for variables

The variables in the three original cohorts are named with a two-letter prefix, e.g. ‘m1’ that identifies the cohort and survey wave.

The letters are y (1973-78 cohort), m (1946-51 cohort), and o (1921-26 cohort)

The 1989-95 cohort, also referred to as the New Young Cohort, or NYC, has been allocated the one-letter abbreviation ‘z’ because it follows on from the first young cohort, which used ‘y’. However, the variable names in the 1989-95 cohort data do not use the prefixes that are used in the other cohorts.

Associated documentation files

Label files allocate meanings to variables. E.g., m1q1='How is your health now?'
Format files allocate meanings to the values of variables. E.g., 1=very good, 2=good etc.

Other Data Files

As well as the survey datasets, there are some supplementary datasets that have been created. Information about dates of deaths and withdrawal of participants is available in the participant status file.

The qualitative data recorded on the back page are also available for analyses. For further information, refer to the Qualitative processing protocols at www.alswh.org.au/how-to-access-the-data/alswh-data.

Child Dataset

There is a Child dataset for the 1973-78 cohort which contains information on birth delivery and related issues for each child. There is only one Child dataset, and at the time of writing it contains information from Surveys 3 to 7, and it is named WHA34567ychildB. There is one record for each child. The Child dataset will be updated with each new survey that has child related information.

Medications datasets

The fourth survey of the 1921-26 cohort, the fifth and sixth of the 1946-51 cohort and the fifth and sixth of the 1973-78 cohort have data on self-reported medications the respondents are taking. These data are available on separate datasets. Where possible, the medications are given by name and ATC code.

Participant Status and Cause of Death files

For a detailed description of Participant Status and Cause of Death files please see section 8 of the Data Dictionary Supplement page.

Extra resources to support data analysis

The Data Dictionary is a Microsoft Access database that gives a detailed description of the questions used in the survey, their source and how they are used, as well as information on the derived and calculated variables. The Data Dictionary is constantly updated and is available at http://www.alswh.org.au/for-researchers/data/data-dictionary (**The table is over 1,000 pages long so do not try to print it**).

The Data Dictionary Supplement is a series of documents that accompanies the Data Dictionary. The Data Dictionary Supplement contains information about scales and other measures used in the ALSWH surveys. Before using any summary or scale score included in an ALSWH dataset, the appropriate section of the Data Dictionary Supplement should be reviewed. The Data Dictionary and Data Dictionary Supplement are available at: http://www.alswh.org.au/for-researchers/data

Check the survey data books if unsure about response frequencies. Electronic copies of the surveys and data books are available at: http://www.alswh.org.au/for-researchers/data

Notes about analysing the data

In general, it is the responsibility of the analyst to become familiar with and carefully examine all data before proceeding with data analysis.

There are different naming conventions for survey items and derived items. IDalias is a unique de-identified participant number, present in all data files. This participant number can be used to merge data files across surveys. The survey questions and method used in the calculation of the derived variables are listed in the Data Dictionary. A few survey items at Survey 1 (birth date, country of birth, language spoken at home) were removed or aggregated into groups, as these were considered potentially able to make participants identifiable.

It is not recommended to arbitrarily replace missing values with the null value or any other value. Questions involving "mark all that apply" responses have been coded to 0 (no response) or 1 (yes response). In general, a "none of the above" response option was offered at the end of each set of "mark all that apply" questions. If responses to all sections of a specific question were missing, including the null option ("none of the above"), all responses were set to missing.

Notes about specific variables

Scales

Regarding items that form part of a scale, be careful that you do not inappropriately analyse single items from a scale. For example, the 36 items in the SF-36 should not be considered as separate items, other than the first self-rated health item. The Data Dictionary Supplement has details about which scales have been included in the surveys.

Counting symptoms

When looking at symptoms, the general rule is to count the number of women who had the symptom "sometimes" or "often".

Measure of depressive symptoms

The 10-item CES-D scale has an extra item at the end ("I felt terrific") which is not included in the calculation of the CES-D score. The CES-D score is available in the datasets.

Menopause

The menopause status variable was calculated at each survey incorporating previous surveys' information for the 1946-51 cohort during the time the women were experiencing menopause.

Measures of physical activity

the physical activity questions were changed after Survey 1. The new physical activity measures from Survey 2 are not comparable to Survey 1 in longitudinal analysis. Refer to the Data Dictionary Supplement for more information.

Summary variables

There are a few "standard" ways to collapse some of the main categorical variables we collect. For example, education (highest qualification) can be dichotomised as "school only", "post school" or in three categories: "no formal qualifications", "school qualifications", "trade/tertiary qualifications" and so on. There have been several variables created to summarise sets of items in the surveys (eg. the illicit drug use items) and it is important that data analysts become familiar with these new variables (See Data Dictionary Supplement)

Area of residence

The recommended measures are ARIA+, present on all surveys, and Modified Monash Model, MMM, only present on surveys after 2012. ARIA+ is an index of accessibility/remoteness based on the distance to the nearest service centre. The scores range from 0 to 15 and the ABS has defined 5 categories for remoteness: major cities of Australia, inner regional Australia, outer regional Australia, remote, and very remote. Only a few of the study's women live in very remote areas, so the fourth and fifth categories are often grouped together. Aria+ and MMM are recommended over the previously used RRMA area classification. For more information see: http://www.adelaide.edu.au/apmrc/research/projects/category/aria.html

For the Modified Monash Model, see the Data Dictionary Supplement section.

ATSI status

Asked at Survey 1 in all age groups. This variable can be used in statistical models but results should not be reported separately by ATSI status in any reports.
See http://www.alswh.org.au/for-researchers/data/indigenous-policy for more information.

Short Surveys

Shorter questionnaires have been used for some respondents in Women's Health Australia when the women had not responded and was contacted late and offered a short survey to complete. The short surveys were only offered in the second surveys of the 1921-26, 1946-51, and 1973-78 cohorts, and the third survey of the 1946-51.

The short surveys only contained those questions that were considered particularly important. These questions are listed in the Short Surveys document. The researcher can identify which respondent did the short survey because their 'survey' variable will have the value 2 rather than 1. These records will have many variables that are entirely missing; the variables that were not included in the short survey.

For more information about using study data and applying to the Publications, Sub-studies and Analyses Committee for access to the data please refer to the website: www.alswh.org.au/how-to-access-the-data/alswh-data.