The Data

ALSWH datasets are comprised primarily of survey items and derived variables which have been calculated using combinations of survey items. Datasets are compiled and distributed under strict privacy provisions and in accordance with the data needs of specific research questions.

Each ALSWH participant has been allocated 2 unique identifiers (code numbers): the study ID and an IDALIAS (an encryption of study ID). These identifiers are used to control access to personally identifying information such as name, address, and date of birth. Records linking the two identifiers are stored securely as per study protocol. As per recently developed protocols for the use of linked data in health research,[1] datasets that contain only the IDALIAS are considered to be re-identifiable or 'pseudonymised', because individual identifier for each participant has been encrypted.

Each item in the ALSWH datasets is allocated one of 3 access levels. Items allocated to access level A are available only to the ALSWH Data Manager and data management staff, and include the study ID. Level B items are more generally available and are included in datasets with the IDALIAS. These items are available to study staff and approved collaborators. Items from food frequency questionnaires and the nutrient data derived from them are assigned level C access and are made available to approved collaborators who are addressing research question that require these data. Datasets including level C items also use the IDALIAS.

Data are stored and distributed as tab-delimited text files (or as SAS datasets on request) using the following naming conventions.

Dataset names

Dataset names have the format wha_*survey_cohort_access level*. Survey is a number between *1* and *5*, indicating when the survey was administered; for example, 1 for the first survey of each cohort, 3 for the third survey of each cohort. Cohort is abbreviated to 3 characters as *yng* for Younger, *mid* for Mid-age and *old* for Older women. The letter *B* indicates that data in this file are re-identifiable.

Example: The dataset containing the re-identifiable responses to the second survey of the Older cohort is entitled wha2oldb.

Variable names

Names for survey variables have a 2-character prefix specifying the cohort and the survey to which they apply. The first character indicates the age cohort, i.e. *y* for Younger, *m* for Mid-age and *o* for Older women. Survey numbers are the same as for dataset names.

Variables names for survey items concatenate the 2-character prefix and the question number. Names for derived variables concatenate the 2-character prefix and a variable descriptor up to 10 characters.

Example: Two variables within the wha2oldb dataset are o2q18a, which is a survey item, and o2bmi, body mass index, a measure of adiposity calculated from height and weight.

References

1. Kelman CW, Bass AJ, Holman CDJ. Research use of linked health data - a best practice protocol. *Australian and New Zealand Journal Public Health* 2002;26:251-5.