AIHW EO2013-1-7:

The Australian Longitudinal Study on Women's Health linkage to the Australian Cancer Database (25th July 2018)

Overview

This report outlines the methodology and results from a probabilistic linkage of survey data from three Australian Longitudinal Study on Women's Health (ALSWH) birth cohorts (1921-26, 1946-51 and 1973-78) with the Australian Cancer Database (ACD). The probabilistic linkage was carried out by the AIHW Data Linkage Unit.

Description of the data

The cohort data set contains 58,172 records (38,552 women identified by LinkID) with linkage data supplied by the Centre for Longitudinal and Life Course Research from the University of Queensland. Each record contains data on name, sex, date of birth, date of last contact, state, death status, date of death and residential address (including postcode).

The cohort of ALSWH women was linked to the ACD. The ACD includes cancers from 1982 to 2013, excluding New South Wales which only includes cancers up to 2012. The linkage data available in the ACD includes name, sex, date of birth, date of diagnosis, date of death, state and postcode.

Method

Data checking

In the AWLHS cohort data set, no records had missing data for surname, given name 1, given name 2, sex or date of last contact. Only one person did not have a date of birth. No postcode information was available for 2.4% of records (1,415) (postcode was extracted from the residential address field). There was 28 records that had no state information and 395 records in which the state was recorded as 'OS' (overseas).

The original ethics amendment requested permission to link survey data from all four ALSWH cohorts (1921-1926, 1946-1951, 1973-1978 and 1989-1995). However the youngest person within the cohort was born in 1979. Therefore the cohort of women born between 1989-1995 were not provided to AIHW to be linked to the ACD. AIHW made the client aware of this and they explained that the cohort of women born between 1989-1995 were not included due to a lag in approvals.

Data linkage

The data sets were linked using probabilistic linkage algorithms. In probabilistic linkage, the linkage of records in two files is based on the probabilities of agreement and disagreement between linkage variables. Probabilistic linkage allows for variation in reporting by allowing

probabilities of agreement to be less than 1 and probabilities of disagreement to be greater than 0.

The probabilistic linkage procedure involves creating record pairs—one from each data set—by running a series of passes that allow for variation in full name information and demographic data. Each pass consists of deterministic pairwise matching on selected blocking variables and then calculating a comparison weight based on probabilities of agreement and disagreement for the blocking and match variables for each respective match pair in the block. In this way, the linkage process creates record pairs by combining records from one data set with records from another data set based on similarities in characteristics such as surname, given name(s) and day, month and year of birth.

NB: It should be noted that probabilistic linkage does **not** require an exact match between two records for any given variable. For each record pair, a record pair comparison weight is calculated. This is an index of the degree of similarity between records in a given pair. It can also be used to ascertain the extent to which a given record pair is likely to be the same person. A higher comparison weight suggests that a given record pair is more likely to be the same person than a lower comparison weight.

Clerical review—general description

Clerical review is the name given to the process that involves manually examining available linkage data for proposed match pairs and deciding whether to accept or reject the match. Commonly, in name-based matching two weight cut-offs are set, with weights above a first (higher) cut-off limit assumed to indicate a match and weights below a second (lower) cut-off assumed to indicate a non-match. Clerical review is then used to decide the match status of possible match pairs with weights between the two cut-offs; that is for record pairs in the 'grey zone' defined by the two weight cut-offs. Clerical review may be carried out after each pass, or after all passes.

In full clerical review, multiple passes are run starting from a high number of blocking variables to a low number of blocking variables. During the initial passes, a large number of blocking variables are applied to identify as obvious true links. Gradually, the number of blocking variables is reduced to allow for more flexible matching (for example, if there is variation in spelling or reported date of birth) to occur. In this way, links with the strongest evidence (highest weights) are identified easily while also allowing true matches with inconsistent data to be identified.

Generally, most links are identified during the initial passes (up to 90%) and few at the end. In the final stages, records are brought together with the least number of blocking variables possible. These final passes allow us to review if there are any more links that were missed as a result of the constraints imposed by the blocking variables. By doing this, we ensure that the search for links is, as far as possible, exhaustive.

Clerical review process for current project

In the current project, clerical review was carried out after each pass, rather than after all passes. As blocking variables were removed, progressively more clerical review was performed to ensure the identification of matches among record pairs with differing linkage information.

Overall, 27 passes were undertaken to create AWLHS cohort to ACD record pairs. Various combinations of the following variables were used as blocking variables:

- Surname

- Given name(s)
- Sex
- Day, month and year of birth (individual elements)
- Postcode

Variables used to assist in clerical review included:

- Surname
- Given name(s)
- Sex
- Date of birth
- Date of last contact / Date of diagnosis
- Date of death
- Postcode
- State

For example, in the first pass (the most restricted pass), record pairs were created by blocking records according to surname, given name(s), sex, full date of birth and postcode. In subsequent passes, the number of blocking variables was reduced; for example, blocking on surname, given name(s), sex and year of birth in order to allow for variation in day and month of birth.

Because the linkage strategy was a probabilistic process, a small percentage of the identified matches may not be correct, and a small number could have been missed even though full clerical review was carried out. In addition, because people's information can be reported quite differently in different data sets, the clerical review process itself is not 100% accurate. However, it is expected that the numbers of false and missed matches is very small.

Additionally, links were checked for information inconsistencies. For example, the date of death / date of last contact (from the AWLHS cohort) was compared to the date of death / date of diagnosis (from the ACD) and links were manually checked.

Results

Overall, 5,420 (14%) women out of 38,552 ALSWH cohort members were matched to the Australian Cancer Database.