CENTRE FOR
HEALTH RECORD
LINKAGE

EDIT CHECKS TO PERFORM BEFORE ANALYSING YOUR LINKED DATA

Familiarise yourself with the data collections in your study

- Read the data dictionaries and contact the data custodian if you have queries about the data collection.

- Be aware of changes in coding of variables over time, and differences in coding of identical variables across datasets (e.g. a value of ‘1’ for a variable in one dataset may have a different value in another).

- Familiarise yourself with ICD codes if relevant and be aware of changes in codes and number of columns available for coding across years in different datasets.

- Review the literature for validation studies on the data items (including ICD codes) you are using or refer to the CHeReL website for information on validation studies http://www.cherel.org.au/validation-studies.

- Be aware of records that may be missing from linked datasets due to the scope of the data collections. For example:

    - Deaths and hospitalisations that occur outside of NSW for NSW residents are not included in NSW datasets linked by the CHeReL. Complete follow-up data are only available for persons who receive all their hospital care in NSW and, if deceased, where the death was registered in NSW.

    - Admitted patient data is provided by year of hospital separation, whereas the Perinatal Data Collection is provided by year of birth of the baby. Some babies may be born in one year and discharged from hospital in the following year.

Data quality and data linkage

- Data collected for administrative purposes are not perfect. The data are collected and entered by humans, often in difficult circumstances (e.g. busy Emergency Departments). A simple key stroke error can lead to an incorrect year of birth, date of admission or variable value. While the CHeReL performs extensive checks on records in the MLK in order to maximise linkage rates (see http://www.cherel.org.au/quality-assurance), records in the source datasets are not altered.

- Be aware of inconsistencies in values for persons within and across different datasets. For example, the recording of whether a person is Aboriginal or Torres Strait Islander is generally based on self-report, and can therefore vary from record to record within and across datasets according to whether the person is asked the question by busy administrative staff, and chooses to identify him- or herself as Aboriginal or Torres Strait Islander on a particular occasion.

- A small proportion of linkage errors are expected. The CHeReL uses probabilistic linkage methods in which error rates are generally around 5 per 1,000. A false positive rate of 5 per 1000 means that in a dataset of 100,000 persons (100,000 Project Person Numbers (PPNs)) you can expect the records of around 500 PPNs to contain linkage errors. Inconsistent records in your dataset may, therefore, be actual false positive links which you may wish to exclude from your analysis. For example a combination of clinical variables (e.g. diagnoses) may indicate that it is unlikely the records belong to the same person. As the CHeReL carries out linkage using demographic variables, inconsistencies in clinical variables are not checked. However, also be aware that these records may in fact be correct links, which the CHeReL has been able to determine through access to the full range of demographic information including first name, surname, date of birth, sex and address. Errors in the variables in the source dataset, eg. an error in date of birth, can make linked records *appear* to be false positives.

Perform basic frequency analyses in individual datasets before merging data to check frequencies and summary statistics.

These might include checking:

- that the records you have received from the data custodian match those reported in the CHeReL methods and measures document;
- for duplicate records prior to and after merging records;
- the total number of cases is close to that expected;
- if annual incidence looks correct;
- whether sex distribution seems appropriate;
- the dataset contains the correct years of data;
- the data do not contain records outside years requested (note however that if hospital episodes are selected based on separation date, the admission date could be outside the period of interest); and
- frequencies against other data sources.

Logic checks

These might include checking for:

- duplicate records;
- ‘missing’ records (e.g. a hospital discharge code of ‘Death’ with no matching entry in the death data);
- unlikely values (e.g. gestational age of 65 weeks, person aged 130 years);
- illogical groupings (e.g. female with prostate cancer, a pregnant male, female giving birth within 3 months of previous delivery);
- illogical sequence of events or variables, such as:
    - DOB<date of diagnosis/service event<DOD

- Mothers DOB<babies DOB

- Age if still living at end of follow-up – a 120 year old individual suggests that the death record is missing or the person moved from NSW

- Date of admission is after the date of death (note that the date of separation may be after the date of death, for example where the person dies before midnight and continues to occupy the hospital bed after midnight).

Dealing with Errors

- Exclude patients or censor records

- Consult a statistician

- Post experiences on the Health Data Linkage forum https://cgi.cse.unsw.edu.au/~hdl/forums/viewforum.php?f=5

- Contact the CHeReL at cherel@cancerinstitute.org.au if you suspect there are systematic linkage errors in more than 5 per 1,000 PPNs.