Government of **Western Australia**
Department of **Health**

Data linkage – making the right connections

Measures of linkage quality

The quality of data linkage processes can be assessed in a number of ways. The most common approaches are by measuring levels of:

- completeness – records in a data collection that have been linked to at least one other record in the linkage system
- false positives – links created in the linkage system that are incorrect
- false negatives – links that were not created in the linkage system but should have been.

Interpreting measures of linkage quality

Measuring completeness is relatively straightforward and the Data Linkage Branch (DLB) provides a completeness metric (expressed as a percentage) for all project-based linkage datasets. The metric can also be provided for other datasets upon request.

As completeness is not always a good indicator of linkage quality, caution is needed when interpreting this measure. If a dataset contained many recent arrivals to Western Australia, for example, it might be legitimate that no other record could be found in the Western Australian Data Linkage System (WADLS) to which these new records could be linked. It would therefore be inappropriate to describe this as an instance of poor quality linkage.

False positives and false negatives are both difficult to identify and measure. Estimates of both measures can be obtained using “mock” data which – theoretically – knows the “right answers” about linkage quality in advance. But it also relies on the mock data realistically mimicking the full range of complexity and data-quality issues found in “real” data. This is impractical and unreliable (especially when data quality in administrative datasets is unknown or shifts unpredictably). As yet, there is no accepted practice for verifying the validity of mock data.

False positives are mistaken links that have not been detected through the linkage process or any related quality-control measures. They are difficult to identify and to find them systematically would require a comprehensive manual review of all links made in the WADLS. As the WADLS is a very large system (~4.2 million chains comprising over 88 million records and spanning more than 40 data collections), reviewing a representative sample would be a massive and/or unfeasible undertaking.

Similarly, false negatives are difficult to measure. These are “unknown unknowns” that were not identified at any stage in the linkage process. A truly comprehensive search would require a manual comparison of every potential pairing of records across all datasets in the WADLS, which would be impractical. Also, not all false negatives are evident from the available data. Data collections have varying degrees of quality, and records sometimes lack the key information necessary to create a single, unambiguous link.

How the Data Linkage Branch maximises linkage quality

The Data Linkage Branch’s linkage process is multi-faceted and includes numerous automated and manual sub-processes designed specifically to reduce the likelihood of errors occurring.

Among these are:

- DLB clerical review – all pairwise comparisons achieve a weight (or score) indicating the likelihood of a match. Pairs of records with high weights are automatically accepted; pairs with low weights are automatically discarded. These weights are determined by Linkage Officers through data evaluation and testing. Some pairs fall into a “grey area” which means they are too strong to discard but not strong enough to match automatically. These pairs are evaluated manually by Linkage Officers. Currently, clerical review accounts for about five per cent of links created in the WADLS. Importantly, clerical review makes it possible to find links or errors that would not have been found using an entirely automated process.

- DLB chain sampling – part of the clerical review process, chain sampling involves reviewing the pair of records under consideration plus all other records to which they have linked. This informs Linkage Officers’ decisions by providing important additional context.

- Data intimacy – Linkage Officers spend time viewing the data closely. This increases familiarity with the data, improving understanding of the wide variety of datasets linked by the DLB, leading to better decision making and linkage strategies.

- DLB’s “weight overrides” – are factored into the linkage processes, whereby particular field combinations can lead a pair of records to be automatically matched, flagged for review or discarded. This fine-tuned control reduces the likelihood of missed or incorrect links.

- DLB’s “no-links” system – ensures that any previously-linked records that have been split, cannot be re-linked without confirmation from a Linkage Officer. This reduces the likelihood of errors being corrected and then inadvertently repeated.

- DLB’s “extra links” system – ensures that records from datasets with restricted visibility (such as sensitive reproductive technology data) will always ‘piggyback’ with the linked record that has the nearest matching demographic information (name, date of birth, etc). If the best-matching record moves to a new chain, the restricted record follows automatically.

- DLB duplicate checking – occurs when one record or chain links to multiple others. Depending on the dataset, this may be treated with suspicion and undergo additional checking. For example, it should be impossible for one inpatient separation to link to multiple birth registrations; further investigation occurs in these instances.

- DLB dynamic “link flagging” system – identifies unlikely or impossible events, record combinations or sequencing, based on the attributes of all records being brought together by a proposed link. Examples include:
    - identification of events that take place before birth or after death
    - multiple instances of a record type that should be unique (e.g.electoral roll registrations, drivers licences or deaths); or
    - car crashes where the driver is less than sixteen years of age.

  The DLB’s “link flagging” system includes 25 different checks and can be adapted dynamically to incorporate new checks or modify existing ones. Most incorrect links will trigger at least one of these flags due to suspicious or incompatible record combinations, making link flagging one of the Branch’s most powerful tools for reducing errors.

- DLB data quality statements – at the conclusion of each linkage process, the DLB produces data-quality statements that include the number of records in a dataset that were not linked, and the number of these that could not reasonably be classed as “linkable” (due to a lack of necessary information). DLB acknowledges there may be legitimate, yet unapparent reasons, for a record failing to link (e.g. clerical errors when the data was first recorded, or genuine singleton records that have no other records to which they can link). Despite this, data quality statements provide a rough estimate of the rate of false negatives. For most of the DLB’s routine linkages, this figure is typically around two to three per cent.

Given these robust checking procedures, the DLB is confident that the rate of errors in the WADLS is very low. However, it should be noted that links are created and modified every day. Each time the DLB acquires new or corrected data, it can result in records moving between chains, including one chain being subsumed by another. The result is that data takes some time to “settle” following initial linkage. Records that have been in the WADLS for less than a year have a one to two per cent likelihood of shifting chains, and this likelihood decreases the longer the record remains in the system. For older data (10+ years), the likelihood of shifting chains is ~0.05 per cent.

The WA Data Linkage Branch prides itself on producing linkages of the highest quality. For more information, please contact the Manager of Data Linkage Systems:

Tom Eitelhuber
E: tom.eitelhuber@health.wa.gov.au
T: (08) 9222-2371

This document can be made available in alternative formats on request for a person with a disability.