SA·NT DataLink

Master Linkage File – Linkage Quality Assurance Report

- Standard Quality Assurance Checks
- Empirical Quality Assessment Stage 1. Review July 2020

Issued: 12 October 2020
Version: 1.0

Table of Contents

1. Scope 1

2. Linkage Overview 1

3. Standard Quality Assurance Checks performed 3

3.1. Check for false negatives 3

3.2. Check for false positives 3

4. Empirical Quality Assessment completed July 2020 3

4.1. Overview 3

4.2. Methodology – Detailed Description 4

4.3. Results 4

4.4. Conclusion 5

1. Scope

This document has been prepared by SA NT DataLink to provide an overview of quality checks routinely undertaken for project specific linkages, and to report empirical outcomes from the review of quality of the Master Linkage File undertaken in July 2020.

2. Linkage Overview

The following linkage quality assurance processes are undertaken to support the highest quality matching of records.

Master Linkage File

The Master Linkage File (MLF) holds all linked records containing people's demographic details. The MLF is updated when new records are added, initially using a deterministic matching approach, which in some cases is sufficient with a high quality match, or where necessary a probabilistic linkage and, the use of a clerical review process for data being added, as well as reviewing project specific extracts.

Preliminary cleaning processes

When datasets are received they are reviewed for completeness, standardised and then loaded into the database.

Generation of derived variables

When a project data linkage request is received, SA NT DataLink also interrogates other datasets in the MLF outside of those specifically requested, to determine if there exist probable individuals that fall within the parameters of the research project. It may for example, search the MLF for previous surnames that belong to a person in the requested dataset and, where relevant, add those records from the MLF to an existing group or create a new group for the project. This process may increase the number of probable individuals identified as relevant to the project.

Deterministic Linkage

Deterministic linkages between records belonging to the same individual are created based on the matching of several identifying variables that have a high level of reliability. Apart from the higher reliability of deterministic linkage, two other benefits are:

- Reduction of the number of duplicated records requiring to be linked probabilistically; and
- Since only one of the number of records determined as belonging to a person needs to be used for linkage, it reduces the demand on the linkage process that would arise from having to link with all of the records of that person.

For example, multiple records within a dataset and/or across different datasets are established as belonging to the same person by having exact matches across eleven variables: *Client ID, Name_Last, Name_First, Name_Others, Date_Birth, Sex, Date_Death, Birth_Weight, Street Address, Suburb* and *Postcode*.

Exact matches for the above variables result in a determination as to whether the record belongs to an existing group of records for that person, or whether a potential new group for that person should be created, or it may remain unmatched. With ten years of operating the SA NT DataLink system, the linking of records to existing or newly created groups is an iterative and ongoing process.

As records are matched ‘exactly’ to a group, the remaining unmatched records are compared and, based on the matching of the variables with a high degree of reliability; using a deterministic rule. The records are identified as belonging to existing or new groups. This process is repeated until exact matching is no longer deemed of high reliable and complete, and are returned for probabilistic linkage.

De-duplication

De-duplication is reliant on the deterministic linkage process and undertaken to decrease the number of records involved in the forthcoming probabilistic linkage. Since all the records out of a group of records are determined as belonging to the same person, it requires that only one variation of the records be taken forward for the proposed linkage.

Probabilistic Linkage

Probabilistic linkage enables the records of persons not linked deterministically to be systematically treated and analysed as likely or unlikely matches depending on the weighted score comparing the linkage variables (that is, more probably):

1. to belong to an existing group of records for the same person; or
2. to be one probable individual so a new group is to be created; or
3. left as a single unmatched record.

Several linkage variables are used to compare records, with each of the linkage variables having a defined weighting. Most usually, these variables are names (family/given/previous names/alias), date of birth, sex, address. Weights are higher for variables with more specificity (such as family name or date of birth) and lower for variables with less specificity for ascertaining individuals (such as sex or given names). The level of agreement between records is determined by the total weight score for every pair of records.

Potential record pairs are classified into three categories: *matches*, *potential matches and non-matches*. The parameters for determining each of these categories are based on the total of weight score attributed to each of the linkage variables, with threshold scores established for each of the parameters. These determine if:

1. Records are considered as a match. That is, as likely belonging to the same person and therefore will be linked.
2. Records are considered potential matches. Each of these group of records are manually reviewed (Clerical Review). The clerical review process decides the final match status (match or un-match) of the records that fall outside of the thresholds.
3. Records remain as unmatched and therefore cannot be linked to another record. That is, a singleton.

3. Standard Quality Assurance Checks performed

The linkage process for adding new datasets or updates to existing datasets is a combination of deterministic and probabilistic methodologies, with clerical review applied to verify possible links. The review of linked records and the standard set of business rules to identify linkage errors provides a high confidence on there being a small percentage of errors in the project outputs. These errors can be false positives (records that have been brought together when they really belong to different individuals) and false negatives (records that have not been brought together when they really belong to the same individual).

The following business rules are applied to identify suspicious records for clerical review.

3.1. Check for false negatives

Criteria for pulling up records to review:

(1) All unmatched study cohort participants from the study cohort are clerically reviewed to identify potential additional links.

(2) Records with missing dates of birth are targeted to try and find links based on other criteria

(3) Unique agency record identifiers that belong to more than one Project Specific Linkage Key (PSLK).

3.2. Check for false positives

The following section outlines the business rules that were used to locate some, if not all, potential false links. Business rules or logic checks used include:

(1) Groups that contain more than one Death Register record.

(2) Groups that contain more than one Perinatal or Birth Register record.

(3) Groups that have records indicating a hospital admission or an emergency presentation after the date of death.

(4) Groups that contain more than one unique agency record identifier.

4. Empirical Quality Assessment completed July 2020

4.1. Overview

In July 2020 SA NT DataLink completed a quality assurance review to assess the overall linkage quality of SA NT Datalink's Master Linkage File (MLF).

Traditionally, linkage quality is assessed by counting false matches and missed matches amongst links produced during linkage. These numbers are then used to produce metrics such as match rate, sensitivity, positive predictive value. F-measure, etc.

While such approach is suitable for an individual (or project-specific) linkage, it cannot be used to assess the overall quality of an enduring MLF, where links (deterministic, probabilistic and clerical) are not kept.

Furthermore, the whole might be greater than the sum of its parts - if we want to find out how well the records are grouped, we need to analyse the groups themselves.

4.2. Methodology – Detailed Description

Instead of analysing individual links, the focus is on assessing the groups, looking for missed links and records that don't belong. Groups represents units of our population of 'probable individuals'.

The process was conducted in two separate analyses and recording the:
1. identification of false linked groups, where a falsely linked group is defined as any group containing at least one record that does not belong to that group
2. identifying incomplete groups, where an incomplete group is defined as any group not containing records that should be part of that group (those records must already be in the MLF and the decision whether they belong to the group is left to experienced clerical reviewers).

The objective is to estimate an upper bound on the rate of incorrectly allocated records using a one-sample binomial proportion test. However a priori there was little idea of the true error rate. Therefore, it was decided to perform the review in two stages: an initial stage 1. assessment and a more extensive review to follow.

To be able to assess the proportion of incomplete groups, a probabilistic linkage was performed. Records belonging to sampled groups were matched against all remaining records in the MLF (meeting inclusion criteria). Matches with cumulative weight above the threshold for clerical review were included in the review process.

A random sample of required size (using SQL built-in sampling function) was selected from Master Linkage Keys (MLKs) meeting the inclusion criteria (loaded into MLF, included in at least one linkage and/or review).

Two experienced clerical reviewers were selected to independently review the sampled groups. They were tasked to classify the groups as 'correct' or 'falsely linked' when looking for falsely linked groups and as 'correct' or 'incomplete' when looking for incomplete groups. Additionally, they were asked to flag each group where a decision can't be made using only the information contained in the records themselves. The results were compared, and all disagreements and flagged groups were reviewed by two additional, independent clerical review officers. All available information, including familial links was used to make the final decision about each group in question.

4.3. Results

The sample size for the Stage 1. review was calculated using one of many freely available online calculators (https://stattools.qacrab.org/R/One_Arm_Binomial.html).

Falsely linked groups

Sample size calculation for the analysis of falsely linked groups was performed to reject an error rate of 0.5% (5 in 1000 groups) in favour of an alternative of 0.1% (1 in 1000 groups). With N=1398 there is 90% power to reject the null hypothesis in favour of the alternative (1-sided alpha=0.05). Due to logistical constraints the Stage 1. sample size was eventually set at N=1200, of which 5 errors were identified, ie 0.4%, with the one-sided exact binomial 95% confidence interval providing an upper bound on the error rate of 0.9% (i.e. 9 in 1000 groups).

These results informed the sample size calculation for the extensive review. With N=3700 there is 80% power to reject a null hypothesis of 0.7% error rate in favour of an alternative error rate of 0.4% in a one-sample binomial test (1-sided alpha=0.05).

Incomplete groups

Sample size calculation for the analysis of incomplete groups was performed to reject an error rate of 2.5% (1 in 40 groups) in favour of an alternative of 1% (1 in 100 groups). With N=624 there is 90% power to reject the null hypothesis in favour of the alternative (1-sided alpha=0.05). The Stage 1. sample size was eventually set at N=600, of which 5 errors were identified, ie 0.8%, with the one-sided exact binomial 95% confidence interval providing an upper bound on the error rate of 1.7% (i.e. 17 in 1000 groups).

These results informed the sample size calculation for the extensive review. With N=1400 there is 80% power to reject a null hypothesis of 1.5% error rate in favour of an alternative error rate of 0.8% in a one-sample binomial test (1-sided alpha=0.05).

The extensive review (Stage 2.) is currently in progress.

4.4. Conclusion

The July 2020 Empirical Linkage Quality Assessment Review (Stage 1.) conducted on the Master Linkage File concluded that the quality matching for the whole of the Master Linkage File (SA and NT) is at 99.6% accuracy (0.4% false positive rate) with the upper bound of the 95% confidence interval error rate at 0.9%. While false negatives or missed links rate is at 0.8% with the upper bound on the error rate at 1.7%