

## QUALITY ASSURANCE IN RECORD LINKAGE

20 December 2009

### 1. Introduction

This document describes:

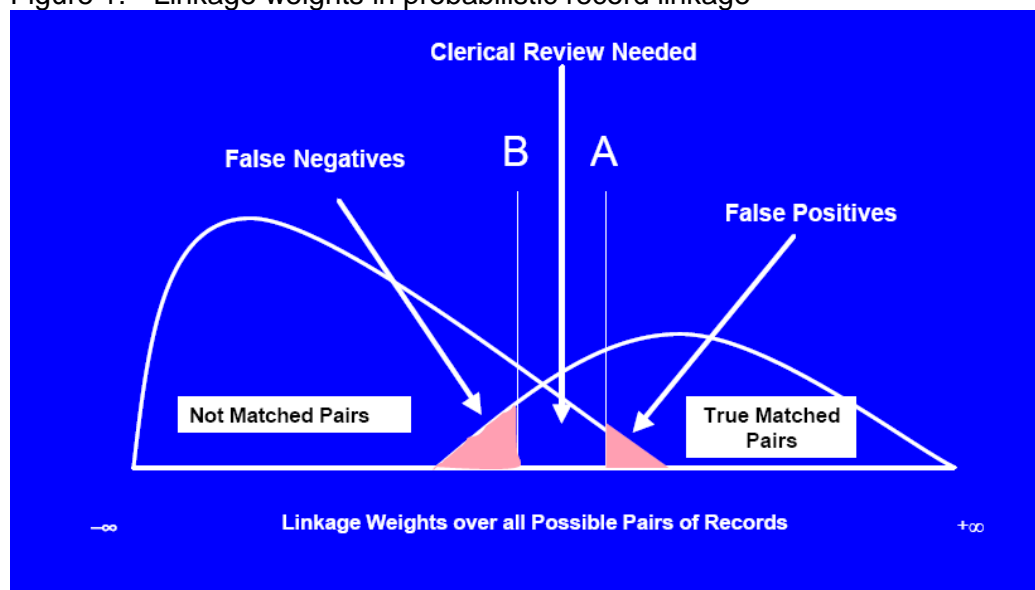
- quality assurance procedures for record linkage projects
- quality assurance procedures for maintenance of the Master Linkage Key
- quality assurance results from the review of the Master Linkage Key (MLK) in December 2009.

### 2. Background

In order to understand the quality of the record linkage at the CHReL, we need to start with Figure 1. Probabilistic record linkage software works by assigning a 'linkage weight' to pairs of records. For example records that match perfectly or nearly perfectly on first name, surname, date of birth and address have a high linkage weight, and records that match only on date of birth have a low linkage weight. Thus if the linkage weight is high it is likely that the records truly match, and if the linkage weight is low it is likely that the records are not truly a match.

However there are pairs of matched records where the linkage weights are neither high nor low, but somewhere in the middle. So how do we decide if they are true matches or not?

Figure 1: Linkage weights in probabilistic record linkage



We could choose the middle linkage weight as a cut-off and arbitrarily say that all pairs of records with linkage weights above the cut-off are 'true' matches, and all pairs of records with linkage weights below the cut-off are 'false' matches.

Unfortunately, this will result in some false matches with linkage weights above the upper cut-off being included with the true matches, and some true matches with linkage weights below the lower cut-off being lost.

At the CHeReL we choose to have two cut-offs:

- an upper cut-off where all pairs of records with linkage weights above the cut-off are designated as 'true' matches;
- a lower cut-off where all pairs of records with linkage weights below the cut-off are designated as 'false' matches; and
- the pairs of records with linkage weights between the upper and lower cut-offs are checked by hand. This is called clerical review (see 3.4 for details).

We aim to adjust the upper and lower cut-offs so that there are:

- no more than 5/1,000 false positive matches above the upper cut-off
- no more than 5/1,000 true positive matches below the lower cut-off (also referred to as false negatives)
- at the same time keeping the amount of clerical review of pairs of records with linkage weights between the upper and lower cut-offs at a manageable level.

Where a linkage project involves records from the MLK, information is collected on whether false positive links relate to records already included in the Master Linkage Key or to the new records being linked to the MLK.

The record linkage software that is used by the CHeReL is ChoiceMaker. ChoiceMaker converts linkage weights to probabilities in the range of 0 to 1, with 0 representing a definite non-match and 1 representing a definite match.

### 3. Quality assurance procedure for record linkage projects

#### 3.1 Set default cut-offs

We start each linkage by setting default cut-offs as follows:

Upper cut-off  $p = 0.75$

Lower cut-off  $p = 0.25$

#### 3.2 Check the upper cut-off

The aim of adjusting the upper cut-off is to minimise the number of false positive matches that lie above the upper cut-off.

A random sample of 1,000 groups of matched records with probabilities that lie above the upper cut-off are reviewed by hand. If the false positive rate is above 5/1,000 the upper cut-off is raised to force these matches into the clerical review area. If there are no false positives, the upper cut-off is lowered to try to reduce the burden of clerical review.

False positive rate	Action on upper cut-off
>5/1,000	↑
0	↓

Once a new cut-off is selected, the linkage is run again and a new random sample of 1,000 groups of matched records that lie above the upper cut-off are reviewed by hand.

The process is repeated until the false positive rate is in the range of 0-5/1,000.



### 3.3 Check the lower cut-off

The aim of adjusting the lower cut-off is to minimise the number of true positive matches that lie below the lower cut-off, because these matches will be lost. We refer to true links that are lost as 'false negative' links.

We review groups of records with probabilities that are close to the lower cut-off. If there are no true matches, then we raise the lower cut-off to reduce the burden of clerical review. If there are true matches close to the lower cut-off we lower the cut-off to try and pick up any true matches that might be lying below the lower cut-off.

A new lower cut-off is selected, the linkage is repeated and groups of records with probabilities that are close to the lower cut-off are reviewed again.

The process is repeated until the false negative rate is no more than 5/1,000.

True positive matches close to lower cut-off	Action on upper cut-off
No	
Yes	

### 3.4 Clerical review of uncertain matches

Groups of linked records with probabilities that lie between the upper and lower cut-offs are reviewed by the CHeReL Record Linkage Officers (RLOs). The RLO compares the records in each group across the full range of available information including first name, surname, date of birth, sex, and address, and decides which records in the group are matches and should stay together.

### 3.5 Quality assurance of Record Linkage Officer (RLO) clerical reviews

Once clerical review of uncertain matches is complete, clerical review is carried out a second time on a random selection of 5% of groups of records that have been reviewed by each RLO. This checking is carried out either by one of the database managers or an experienced RLO. Two types of errors are possible: a 'critical' error, where a clear mistake has been made; or a 'non-critical' error, where the match/non-match decision is very close. 'Non-critical' errors are generally caused by missing information on one or both records.

The consequences of errors are:

- if one 'critical error', all work of RLO for the project is checked
- if 'non-critical error' rate < 2.5%, work accepted
- if 'non-critical error' rate > 2.5%, 50% work for the project is reviewed:
  - if 'non-critical error' < 2.5%, work accepted
  - if 'non-critical error' > 2.5%, 100% work for the project is reviewed

## 4. **Quality assurance on the MLK**

Quality assurance is carried out on the MLK once a year. Quality assurance aims to detect and correct incorrect (false positive) links. A new process was added in 2009 to detect missed links in the MLK. Each person in the MLK has a Person Identifier (Person ID) which is used exclusively by the CHeReL and is not released outside the CHeReL.

#### 4.1 Correction of false positive links in MLK

The quality assurance checks will grow over time. A list of the current checks are:

1. Hospital admission date > date of death
2. Date of birth and first 4 letters of first name and first 4 letters of surname differ
3. Date of birth differs by more than 5 years and first 4 letters of address differ
4. Year of birth differs by more than 10 years
5. Males where date of birth and surname differ
6. Both surname and given name differ (excludes birth records)
7. Hospital admission date is prior to birth date
8. Dates of death differ by 2 or more days
9. Person ID contains Multiple death records and (death month+year or sex or given names differ)
10. Compare Given Names and surname (first 3 characters of given name and first 4 characters of surname after stripping spaces and non-alpha characters).
11. Person ID contains a MDC Mother record and a MDC Baby record or RBDM birth record (This Check is only temporary as MLK only contains 15 years of history).
12. Person ID contains 45 and Up record and a MDC baby record.
13. Person ID has different post code, Date of birth, and (stripped) first 4 characters of given name.
14. Person ID contains MDC baby record and RBDM birth record and Mothers name does not match.
15. Person ID has Death Date before MDC mother baby's date of birth.
16. Person ID contains MDC baby record with another record where age greater than 14 (This Check is only temporary as MLK only contains 15 years of history).
17. Person ID contains more than 1 MDC baby record (checking for twins matched together).
18. Person IDs which contain more than 500 records.
19. Person IDs that contain blank given names, surname and address.
20. Person IDs that have been identified as incorrect since last Key QA process (from Linkage cut-off checking and extracts for study projects)

Any pairs/groups of records that fulfil any of the above criteria undergo further clerical review.

The results of the quality assurance on the MLK that was carried out in December 2009 is shown in Table 2. Prior to corrections being carried out 0.512% of people represented in the MLK had records that failed one or more checks. There were 0.186% of people represented in the MLK whose linked records had been reviewed in 2008 and found to be correct; these were excluded from clerical review in 2009. This left linked records for 0.325% of people represented in the MLK to undergo clerical review. After false positive links were corrected by clerical review process 0.461% of people represented in the MLK had records that failed one or more checks. In these 0.461% of cases the links were correct, but the information on the record that was used in the check was incorrect.

Inconsistent information in the MLK may therefore be due to false positive links or incorrect information being provided from the source database. The overall percentage of records and persons affected is very small. If researchers find inconsistent information in linked records, we recommend that these records be excluded from the analysis and the number and proportion of records excluded be reported in the methods section of the study report.

At the conclusion of the QA process records for 1,000 persons were randomly sampled from the rebuilt Master Linkage Key and the false positive rate found was 3/1000.

#### 4.2 Corrections of missed links in Master Linkage Key

Possible missed links were identified using a statistical linkage key (SLK). The SLK581 is the concatenation of the 2nd, 3rd and 5th letters of the family name, the 2nd and 3rd letters of the given name, date of birth as a character string of the form *ddmmyyyy*, followed by the character '1' for male and '2' for female. Non-alphabetic letters in names are excluded (for example, hyphens and apostrophes), and where a name contains insufficient letters the character '2' is used as a place marker for absent key letters. The character '9' is used for any other missing data so that the linkage key always has a length of 14 characters. Records with incomplete personal identifiers were excluded from the above selection.

A deterministic linkage of remaining records in the Master Linkage Key was carried out using the SLK581. All new links created were subject to clerical review.

There were 166,936 Person IDs in the MLK that were identified as being possible matches with another Person ID, creating a possible 82,697 combined Person IDs. For the 166,936 Person IDs, 108,104 were matched together without clerical review based on exact match of surname, first 3 characters of given names, sex, and date of birth for Person IDs where the surname was not in the 40 most common surnames in NSW. This produced 53,609 combined Person IDs. The remaining 58,832 Person IDs were subject to clerical review and 12,875 combined Person IDs were confirmed. In total 166,936 Person IDs were reduced to 66,484 new combined Person IDs.

**Table 2: Results of correction of false positive links, December 2009**

Test No.	Test description	Pre correction				Post correction			
		Person IDs in MLK 2009	Person IDs in MLK 2008 checked and correct	Person IDs in MLK 2009 remaining for review		Person IDs in MLK post correction		Person IDs in MLK 2009 corrected	
		No.	No.	No.	%	No.	%	No.	%
1	admission date > date of death	1,913	668	1,245	0.017%	1,057	0.024%	188	15.100%
2	date of birth + 4X4 initials differ	65	10	55	0.001%	24	0.000%	31	56.364%
3	date of birth > 5 years + address differs	14	1	13	0.000%	0	0.000%	13	100.000%
4	year of birth > 10 years	12,174	1,332	10,842	0.152%	10,111	0.160%	731	6.742%
5	date of birth and surname differ (sex=M)	2,106	638	1,468	0.021%	1,153	0.025%	315	21.458%
6	surname + given names differ	9,133	4,739	4,394	0.062%	3,313	0.113%	1,081	24.602%
7	hospital admission date < birth date	5,857	2,904	2,953	0.041%	2,165	0.071%	788	26.685%
8	dates of death differ by 2 or more days	1,118	813	305	0.004%	235	0.015%	70	22.951%
9	ID contains multi RBDM Death records	476	150	326	0.005%	232	0.005%	94	28.834%
10	Compare Giv Names (3 chars) + Surname (First 4 Chars)	367	127	240	0.003%	133	0.004%	107	44.583%
11	ID has Mother MDC and (MDC Baby or RBDM birth Records)	1,386	1,064	322	0.005%	296	0.019%	26	8.075%
12	ID has 45andUP rec and baby MDC record	30	4	26	0.000%	26	0.000%	0	0.000%
13	ID has diff Post Code + DOB + diff 4 chars Giv Names	3,507	1,117	2,390	0.033%	2,018	0.044%	372	15.565%
14	ID has MDC baby + RBDM birth + Diff Mother's Name	549	213	336	0.005%	94	0.004%	242	72.024%
15	ID has Date of Death before baby DOB	4	2	2	0.000%	2	0.000%	0	0.000%
16	ID has MDC baby with another record age > 14	9,269	406	8,863	0.124%	8,653	0.127%	210	2.369%
17	ID has more than one MDC baby record	774	236	538	0.008%	82	0.004%	456	84.758%
18	select all MLK IDS where No recs > 500	1,273	242	1,031	0.014%	1,030	0.018%	1	0.097%
19	select IDs where name blank + address is blank	206	1	205	0.003%	0	0.000%	205	100.000%
20	Select specific MLK IDs to review (Problem found reviewing previous linkage upper cut-offs)	20	0	20	0.000%	0	0.000%	20	100.000%
<b>Total MLK Person IDs</b>		36,538	13,292	23,246	0.33%	19,668	0.46%	3,578	15.39%

Total Person IDs in Master Linkage Key at time of Quality Assurance were **7,143,455** persons.