

MLK QUALITY ASSURANCE

30 April 2011

Methods

In 2011 the CHeReL carried out a quality assurance exercise that aimed to detect and correct incorrect (false positive) links and missed links in the Master Linkage Key (MLK). Missed links were processed first so that any linkage errors that were inadvertently introduced could potentially be re-processed and corrected during the checking of false links.

Correction of missed links

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 MLK was carried out using the SLK581. Where the surname was in the top 40 frequently occurring names in the MLK the groups were subject to a manual clerical review process. Where the surname occurred less frequently in the MLK, groups were automatically matched using the SLK-581.

4.2 Correction of false positive links

Checks used in 2011 were:

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 1 year and month or day is also different.
5. Males where date of birth and surname differ
6. Both surname and given name differ (excludes birth PDC and NCIMS 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 PDC Mother record and a PDC Baby record or RBDM birth record
12. Person ID contains 45 and Up record and a PDC baby record.
13. Person ID has different post code, Date of birth, and (stripped) first 4 characters of given name.
14. Person ID contains PDC baby record and RBDM birth record and Mothers name does not match.
15. Person ID has Death Date before PDC mother baby's date of birth.
16. Person ID contains PDC baby record with another record where DOB was before 1 Jan 1994.
17. Person ID contains more than 1 PDC baby record (checking for twins matched together).
18. Person IDs that contain blank given names, surname and address.

MLK IDs that had not changed since December 2009 (the last quality assurance exercise) were excluded.

Results

From the CHeReL (MLK 2011_03) which contained 8,188,866 MLK IDs, 202,001 were identified as being possible matches with another MLK ID. Of these 202,001 MLK IDs, 151,538 were matched without clerical review based on exact match of the SLK-581 where the surname was not in the 40 most common surnames in the MLK. This automatic deterministic matching produced 75,469 combined MLK IDs. The remaining 50,463 MLK IDs were grouped to form 24,721 groups which were subject to manual clerical review process, and there were a total of 34,838 MLK IDs remaining after the manual clerical review process was completed.

In total 202,001 MLK IDs were reduced to 110,307 new combined MLK IDs. These 110,307 changed MLK IDs were then applied to the Master Linkage Key before starting the second phase of the QA process to identify false positive links within MLK IDs.

The results of the false positive links quality assurance on the MLK that was carried out in April 2011 is shown in Table 1. From the 8,188,866 MLK IDs in the MLK at April 2011 there were 2,842,700 new or changed MLK IDs since the last yearly QA (Dec 2009). Applying the above checks produced 25,231 groups (MLK IDs) to be manually clerical reviewed by CHeReL RLO staff. These 25,231 groups contained 32,316 potential errors. During clerical review 36% of these potential errors were confirmed to be linkage errors and fixed.

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 4/1000 MLK IDs.

Conclusion and Recommendation

Inconsistent information in the MLK may 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.

Table 1: Results of correction of false positive links, April 2011

Test No.	Test Description			Post Manual Correction			
		Person IDs in MLK Dec 2009 checked and correct (Unchanged MLK IDs from last QA)	Person IDs selected for QA 2011 (excluding MLK IDs That had not changed since last QA Dec 2009)	MLK IDs in MLK 2011 post correction (Still triggering test to fire)		MLK IDs in MLK 2011 where test no longer fires (error was corrected)	
		No.	No.	No.	% of Total MLKs	No.	% of errors found
1	admission date > date of death on any rec source	1,057	2,570	1,430	0.017%	1,140	44.358%
2	date of birth 4 chars surname + 4 chars givnames differs	24	99	62	0.001%	37	37.374%
3	date of birth > 5 years + address differs	0	5	1	0.000%	4	80.000%
4	date of birth differs > 1 year + month or day not same	10,111	3,627	3,030	0.037%	597	16.460%
5	date of birth and surname differ and sex is male	1,153	2,837	2,515	0.031%	322	11.350%
6	surname + given names differ + (not baby PDC or NCIMS recs)	3,313	6,528	2,903	0.035%	3,625	55.530%
7	hospital admission date < birth date	2,165	4,091	3,954	0.048%	137	3.349%
8	dates of death differ by 2 or more days	235	667	290	0.004%	377	56.522%
9	ID contains multi RBDM Death records	232	415	42	0.001%	373	89.880%
10	Compare Giv Names (3 chars) + Surname (First 4 Chars)	133	782	387	0.005%	395	50.512%
11	ID has Mother PDC and (PDC Baby or RBDM birth Records) (Temp Query will need to change when MLK matures)	296	1,493	1,473	0.018%	20	1.340%
12	ID has 45andUP rec and baby PDC record.	26	26	24	0.000%	2	7.692%

Test No.	Test Description			Post Manual Correction			
		Person IDs in MLK Dec 2009 checked and correct (Unchanged MLK IDs from last QA)	Person IDs selected for QA 2011 (excluding MLK IDs That had not changed since last QA Dec 2009)	MLK IDs in MLK 2011 post correction (Still triggering test to fire)		MLK IDs in MLK 2011 where test no longer fires (error was corrected)	
		No.	No.	No.	% of Total MLKs	No.	% of errors found
13	ID has diff Post Code + DOB + diff 4 chars Giv Names	2,018	4,463	3,807	0.046%	656	14.699%
14	ID has PDC baby + RBDM birth + Diff Mother's Name	94	1,853	204	0.002%	1,649	88.991%
15	ID has Date of Death before baby DOB	2	14	5	0.000%	9	64.286%
16	ID has PDC baby with another record DOB before 1/1/94	8,653	244	138	0.002%	106	43.443%
17	ID has more than one PDC baby record	82	2,412	101	0.001%	2,311	95.813%
18	select IDs where name blank + address is blank	0	190	190	0.002%	0	0.000%
Total MLK IDs		13,292	25,231				

** Note: Total MLK IDs will be less than the sum of all individual tests due to the fact that some MLK IDs fail multiple tests.

Table 2: Summary of Master Linkage Key Person ID movement

Total KEY IDS in MLK:	BEFORE QA	8,188,866
Missed links merged by script (net IDs)		-76,069
Missed links merged by manual review (net IDs)		-15,625
False positive links split manual review (net IDs)		7,998
45 and up duplicates split in new MLK IDs		82
Total KEY IDS in MLK:	AFTER QA	8,105,252

There were also 82 MLK IDs adjusted due to changes in 45 and up records, resulting from confirmation about potential duplicate records.