Master Linkage Key Quality Assurance

16 August 2012

Methods

In 2012 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 manual clerical review process was conducted on MLK IDs that contained records that matched in the above process.

Correction of false positive links

Checks used in 2012 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 and Date of birth was after year 2001 - (Changed 2012 QA)
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. (CHeReL PDC baby records start at Jan 1994) - (Changed 2012 QA)
17. Person ID contains more than 1 PDC baby record (checking for twins matched together).
18. Where number of records in MLK ID greater than 1200. - (New 2012 QA)
19. Person IDs that contains Blank name + blank address - (New 2012 QA)
20. Person IDs that were incorrectly linked using SLK581 as a matching field. (SLK 581 matched but surname or given names do not match) - (NEW 2012 QA)

MLK IDs that had not changed since April 2011 (the last quality assurance exercise) were excluded along with single record MLK IDs.

Results

The CHeReL (MLK 2012_11) contained 9,349,199 MLK IDs.

In the missed links process there were MLK IDs 60,572 IDs identified as being possible matches with another MLK ID. These 60,572 MLK IDs were grouped to form 20,055 groups which were subject to a manual clerical review process which resulted in a total of 20,274 MLK IDs being formed.

Detailed results for the false positive links process are shown in Table 1. From the 9,308,901 MLK IDs in the MLK (after the missed links process) there were 2,492,973 MLK IDs that were new or changed since the previous QA in April 2011 and were therefore eligible for inclusion in the false positive checking process. The checks produced 19,351 groups (MLK IDs) that were subject to a manual clerical review process. These 19,351 groups contained 22,943 potential errors. During clerical review 14.5% 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 3/1000 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, July 2012

Test No.	Test Description	Before Manual Correction	After Manual Correction			
		MLK IDs eligible for the QA process	MLK IDs Uncorrected by manual review		MLK IDs Corrected by manual review	
		No.	No.	% of Total MLKs	No.	% of errors found
1	admission date > date of death on any rec source	**1,620**	**1,473**	0.018%	**147**	9.074%
2	date of birth 4 chars surname + 4 chars givnames differs	**76**	**52**	0.001%	**24**	31.579%
3	date of birth > 5 years + address differs	**9**	**4**	0.000%	**5**	55.556%
4	date of birth differs > 1 year + month or day not same	**3,648**	**3,013**	0.037%	**635**	17.407%
5	date of birth and surname differ and sex is male	**2,509**	**2,361**	0.029%	**148**	5.899%
6	surname + given names differ + (not baby PDC or NCIMS recs)	**3,722**	**3,175**	0.039%	**547**	14.696%
7	hospital admission date < birth date	**3,821**	**3,665**	0.045%	**156**	4.083%
8	dates of death differ by 2 or more days	**381**	**290**	0.004%	**91**	23.885%
9	ID contains multi RBDM Death records	**461**	**346**	0.004%	**115**	24.946%
10	Compare Giv Names (3 chars) + Surname (First 4 Chars)	**471**	**402**	0.005%	**69**	14.650%
11	ID has Mother PDC and (PDC Baby or RBDM birth Records) (Temp Query will need to change when MLK matures)	**13**	**0**	0.000%	**13**	100.000%
12	ID has 45andUP rec and baby PDC record.	**19**	**17**	0.000%	**2**	10.526%

Test No.	Test Description	Before Manual Correction	After Manual Correction			
		MLK IDs eligible for the QA process	MLK IDs Uncorrected by manual review		MLK IDs Corrected by manual review	
		No.	No.	% of Total MLKs	No.	% of errors found
13	ID has diff Post Code + DOB + diff 4 chars Giv Names	**3,930**	**3,645**	0.045%	**285**	7.252%
14	ID has PDC baby + RBDM birth + Diff Mother's Name	**202**	**124**	0.002%	**78**	38.614%
15	ID has Date of Death before baby DOB	**13**	**5**	0.000%	**8**	61.538%
16	ID has PDC baby with another record DOB before 1/1/94	**227**	**94**	0.001%	**133**	58.590%
17	ID has more than one MDC baby record	**212**	**80**	0.001%	**132**	62.264%
18	select all MLK IDS where No recs > 1200 (Not Done Last QA)	**282**	**282**	0.000%	**0**	0.000%
19	select IDs where name blank + address is blank	**190**	**170**	0.002%	**20**	10.526%
20	Select MLK IDs that incorrectly linked by SLK581 (surname given names not match and SLK581 did match)	**1,137**	**0**	0.000%	**1,137**	100.%
Total MLK IDs		**19,351**	**16,541**			

** 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**	**9,349,199**
Missed links merged by manual review (net IDs)		**-40,298**
Total KEY IDS in MLK:	**AFTER MISSED LINKS:**	**9,308,901**
False positive links split manual review (net IDs)		**+2,493**
Total KEY IDS in MLK:	**AFTER YEARLY QA**	**9,311,394**