Completeness and validity of the pseudonymised NHS number in QResearch and utility for data linkage

1 Background

- This article summarises the completeness and validity of the pseudonymised NHS number in GP electronic records based on practices in England contributing to the QResearch database (www.qresearch.org).
- The QResearch database is derived from practices using the EMIS Web and EMIS LV clinical system. Previous analyses have shown that these practices are similar to all EMIS practices and also practices using other clinical systems (such as INPS).
- QResearch does not include any strong patient identifiers but does include a database specific pseudonymised NHS number which has been generated using the Open Pseudonymiser method (www.openpseudonymiser.org) as approved by the Ethics and Confidentiality Committee of the National Information Governance Board and Trent MREC.
- The pseudonymisation method is described in detail at the above link. In summary, in order to generate this pseudonymised, the software concatenates the NHS number with a project specific encrypted password (known as a salt code) and then applies a one way hashing algorithm within the source clinical system. The resulting pseudonym is then a project specific code which does not allow the individual to be identified (protecting confidentiality) but which does allow the data to be linked to other datasets (such as HES, cancer and mortality data) which have been processed in the same way.
- The software also generates a data quality flag which confirms whether the source NHS number has passed the NHS checksum, whether it failed or whether the NHS number was missing. This data quality flag (field name ValidNhsNumber) is recorded for all patients on the database.

2 Methods

- We included all 607 practices in England contributing to the QResearch database on 1st March 2013. We used version 35 of the database (uploaded 6th March 2013).
- We included all men and women who were registered on 1st March 2013.
- We then summarized the numbers of patients with a complete and valid NHS group by the following strata: age, sex, Strategic Health Authority, clinical system type (EMIS LV or EMIS Web).
- Additional analyses of this are available on request (e.g. by age, deprivation, ethnicity).

3 Results

There were 607 practices spread across all ten SHAs in England. There were 5,078,704 currently registered patients. Of these, 5,070,000 (99.83%) had a valid NHS number and 8,704 had a missing NHS number. The table below shows the breakdown by sex, geographical area and clinical system type.

The NHS number is complete and valid in >99.8% of currently registered patients. This is consistent across sex, system type and geographical area.

	valid NHS	number of patients (n=5,078,704)	% complete and accurate
all patients	no	8,704	0.17
	yes	5,070,000	99.83
sex			
women	no	4,232	0.17
women	yes	2,559,330	99.83
men	no	4,472	0.18
men	yes	2,510,670	99.82
EMIS Web			
LV	no	2,916	0.12
LV	yes	2,405,059	99.88
Web	no	5,788	0.22
Web	yes	2,664,941	99.78
East Midlands SHA	no	340	0.07
East Midlands SHA	yes	467,177	99.93
East of England SHA	no	296	0.07
East of England SHA	yes	444,326	99.93
London SHA	no	4,366	0.45
London SHA	yes	965,666	99.55
North East SHA	no	193	0.07
North East SHA	yes	293,791	99.93
North West SHA	no	511	0.08
North West SHA	yes	651,604	99.92
South Central SHA	no	514	0.11
South Central SHA	yes	474,086	99.89
South East Coast SHA	no	446	0.11
South East Coast SHA	yes	388,446	99.89
South West SHA	no	919	0.15
South West SHA	yes	607,926	99.85
West Midlands SHA	no	502	0.11
West Midlands SHA	yes	439,012	99.89
Yorkshire and the Humber SHA	no	617	0.18
Yorkshire and the Humber SHA	yes	337,966	99.82

4 Conclusion

- The NHS number is complete and valid in >99.8% of currently registered patients in general practice computer systems contributing to the QResearch database.
- The extremely high levels of completeness and validity of the NHS number have enabled us to use pseudonymised NHS number as the sole identifier to link QResearch GP data to individual level record data from ONS mortality, cancer records and HES data.
- The correspondence between year of birth and sex on the linked datasets is extremely high which acts as a 'sanity' check on whether the same individual record does appear in each of the four datasets.
- The open pseudonymiser software is now available within EMIS systems (55% of GP practices) and TPP practices (around 20% of practices) and used as a standard approach.
- Therefore similar checks of NHS number completeness can be undertaken on a larger number of practices to verify this (subject to consent from relevant parties).
- However, the evidence so far suggests that the NHS number is likely to be a unique and reliable identifier within GP records and that this can be pseudonymised at source and the pseudonymised NHS number used for data linkage studies.
- Our recommendation is that this approach is tested for other data linkage studies where data has already been collected 'in the clear' to compare its performance against existing methods. This will determine the incremental value of each method and whether linkage on NHS number alone (or pseudonymised NHS number) is sufficient.
- Additional analyses of the completeness of the NHS number on QResearch or the linked data may be requested from Julia.hippisley-cox@nottingham.ac.uk

Acknowlegdments

EMIS practices, EMIS and University of Nottingham for contribution and expertise in developing and supporting QResearch.