Validity and completeness of the NHS Number in primary and secondary care Electronic data in England 1991-2013

Date: 09.06.2013
Version 1.0

Contents

Author Julia Hipisley-Cox, University of Nottingham
Date: 09.06.2013
Version 1.0

1 Executive summary

- This purpose of this article is to document the completeness and validity of the pseudonymised NHS number in six English national data sources 1991-2011.

    1. Cancer Registry data
    2. Deaths Registry data
    3. Accident and Emergency Attendances
    4. Inpatient Hospital Episodes
    5. Outpatient Hospital Episodes
    6. General practices contributing to the QResearch database

- The table below summarizes the figures for each data source over all the years available and also for the latest full year of data available.
- Over 99.5% of ONS Cancer and Deaths Registry records have a complete and valid NHS number.
- For hospital episode statistics, the NHS number is valid and complete for 94% of A&E records, and 98.6% of inpatient and 98.9% of outpatient records.
- For primary care, 99.8% of current patients have a complete and valid NHS number.
- The 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.

Table 1: Summary of completeness and validity of NHS number recording in six primary and secondary care English NHS data sources

	start	end	total records	total records with valid NHS	% records with valid NHS number
All available years					
A&E	01.04.2007	31.08.2012	75,542,582	68,231,107	90.32
Out patients	01.04.2003	31.08.2012	672,277,004	655,674,902	97.53
In patients	01.04.1997	31.08.2012	233,132,981	655,674,902	97.53
Cancer registry	01.01.1990	31.12.2010	6738358	6705806	99.52
Death registry	01.01.1997	31.12.2011	7,809,003	7,800,617	99.89
Last complete year					
A&E	01.04.2011	31.03.2012	17,619,708	7,326,769	94.32
Out patients	01.04.2011	31.03.2012	90,956,844	89,908,970	98.85
In patients	01.04.2011	31.03.2012	18,889,329	18,619,684	98.57
Cancer registry	01.01.2010	31.12.2010	417,389	416,172	99.71
Death registry	01.01.2011	31.12.2011	463,450	463,145	99.93
Primary Care (QResearch)	01.03.2013	01.03.2013	5,078,704	5,070,000	99.83

Author Julia Hipisley-Cox, University of Nottingham
Date: 09.06.2013
Version 1.0

- This approach has significant advantages over existing approaches to data linkage where data is extracted from source systems 'in the clear' since it better protects patient confidentiality.
- The OpenP software has been implemented within EMIS systems (55% of GP practices) and TPP practices (around 20% of practices) and used as a standard approach and two major data controllers of NHS secondary care data (The Health and Social Care Information Centre and the Office of National Statistics).
- This project has demonstrated the utility and scalability of applying pseudonymisation at source to the NHS number recorded in NHS clinical data for purposes of data linkage between primary and secondary care data.
- This approach, which is scalable and cost-effective, is intended to promote the public benefits of data sharing whilst protect patient confidentiality.

2 Background

- This purpose of this article is to document the completeness and validity of the pseudonymised NHS number in six English national data sources 1991-2011.
    - Cancer Registry data
    - Deaths Registry data
    - Accident and Emergency Attendances
    - Inpatient Hospital Episodes
    - Outpatient Hospital Episodes
    - General practices contributing to the QResearch database
- The information in this report may be of use to organisations wishing to undertake data linkage between different data sources using a pseudonymised version of the NHS number.
- Each data source included in this report has been pseudonymised at source by the relevant data controllers using the freely available Open Pseudonymiser Software. (www.openpseudonymiser.org).
- For Cancer and Deaths Registration, the pseudonymisation at source was undertaken by the Office of National Statistics.
- For Accident and Emergency, Hospital Outpatient and Hospital Inpatient episodes, this was done by the Health and Social Care Information Centre.
- For primary care data (QResearch database), this was done by EMIS Ltd in collaboration with the University of Nottingham.
- The purpose of the data processing was to prepare data for the QResearch Multi-Data Source Linkage Project (www.qresearch.org).
- Initially section 251 Support was obtained from the Ethics and Confidentiality Committee of the National Information Governance Board in order to access identifiable data for the purposes of data linkage to the QResearch database.
- However, during discussions, it became apparent the linkage could be done using data pseudonymised at source provided that this was done in an identical way on each of the four data sources to be linked. This would obviate the need for section 251 approval as identifiable data would not be disclosed by the data controllers.
- The approach was reviewed Ethics and Confidentiality Committee of the National Information Governance Board who confirmed that s251 support would no longer be required.
- The proposed method was also approved by the Trent Multi-Centre Ethics Committee, the QResearch Advisory Board and the EMIS National User Group.

3 Methods

- We included five national datasets provided by the Office of National Statistics, the Health and Social Care Information Centre and the QResearch database
    - Cancer Registry data
    - Deaths Registry data
    - Accident and Emergency Attendances
    - Inpatient Hospital Episodes
    - Outpatient Hospital Episodes
    - Primary Care Data (QResearch)
- All records for all patients in England were included for all years for where data was available.
- Data from 607 general practices in England was used for QResearch
- The NHS number in each record was pseudonymised at source using the Open Pseudonymiser Software (referred to as OpenP). This is described in detail at www.openpseudonymiser.org.
- In summary, the OpenP 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 pseudonymised NHS number is then project specific. The pseudonym does not allow the individual to be identified (protecting confidentiality) but does allow the data to be linked to other datasets which have been processed in the same way.
- The OpenP software also rounds dates of birth to year of birth and strips off any identifiers (such as full NHS Number or other strong identifiers). All strong identifiers were removed in the source system by the relevant data controller.
- The OpenP software also generates a data quality flag which flags NHS numbers which have passed the NHS checksum, those which have failed and those where the NHS number is missing.
- This data quality flag is recorded for all patients on each of the data sources and can be used to summarise the validity of the NHS number before data extraction or on receipt of the data. Summary information on this is provided in the log file generate at run time by the OpenP software.
- The source data were then encrypted and transferred securely to the University of Nottingham where each dataset was then examined in detail.
- This report focuses on the completeness and validity of the pseudonymised NHS Number in each dataset prior to data linkage.

4 Results

This section describes the results of each datasets analysed at individual record levels.

4.1 Cancer registry data

- **Data supplier**: Office of National Statistics
- **Time period**: 01 Jan 1991-31 December 2010
- **Coverage**: all patients in England
- **Updates**: annual
- **Description**: Record level data for each cancer registration recorded for patients in England. Key fields include pseudonymised NHS number; data quality flag; sex, year of birth; date of death; site of cancer, type of growth histology; behavior of growth; basis of diagnosis; stage; differentiation; treatment
- **Summary**. 99.52% of cancer registry records have a complete and valid NHS number over the last 20 years. This over 99.7% for the last 10 years.

Table 2 Cancer registry: completeness & validity of NHS number 01Jan1991-31 Dec 2010

calendar year	total cancer records	invalid NHS	Valid NHS number	% valid NHS number
1991	263,620	16,521	247,099	93.73
1992	274,769	2,345	272,424	99.15
1993	271,307	1,408	269,899	99.48
1994	280,859	996	279,863	99.65
1995	287,820	660	287,160	99.77
1996	291,950	660	291,290	99.77
1997	304,913	744	304,169	99.76
1998	313,715	674	313,041	99.79
1999	326,049	762	325,287	99.77
2000	331,035	1,189	329,846	99.64
2001	336,963	817	336,146	99.76
2002	337,183	664	336,519	99.80
2003	344,618	724	343,894	99.79
2004	358,098	638	357,460	99.82
2005	368,719	564	368,155	99.85
2006	381,774	505	381,269	99.87
2007	395,741	468	395,273	99.88
2008	421,749	442	421,307	99.90
2009	430,087	554	429,533	99.87
2010	417,389	1,217	416,172	99.71
1991-2010	**6738358**	**32552**	**6705806**	**99.52**

Author Julia Hipisley-Cox, University of Nottingham
Date: 09.06.2013
Version 1.0

4.2 Deaths Registry Data

- **Data supplier**: Office of National Statistics
- **Time Period**: 01.01.1997 to 31.12.2011
- **Coverage**: all patients in England
- **Updates**: annual
- **Description**: Record level data for each death. Pseudonymised NHS numbers; NHS number data quality flag, year of birth, date of death, cause of death
- **Summary**: 99.94% of death registry records have complete and valid NHS numbers. This has been over 99.87% since 1998.

Table 3 Mortality data: Completeness and validity of NHS number 01 Jan 1998 to 31 Dec 2011

	Patient records	invalid NHS	missing NHS number	valid NHS number	% valid NHS
1997	556,134	104	3,860	552,155	99.28
1998	555,995	87	64	555,826	99.97
1999	557,228	109	60	557,048	99.97
2000	536,839	69	77	536,679	99.97
2001	531,740	76	126	531,514	99.96
2002	535,311	72	341	534,892	99.92
2003	540,265	176	20	540,068	99.96
2004	514,759	74	180	514,503	99.95
2005	515,089	51	178	514,858	99.96
2006	504,300	48	384	503,867	99.91
2007	505,879	61	582	505,235	99.87
2008	508,907	56	312	508,539	99.93
2009	490,198	82	258	489,858	99.93
2010	492,909	98	381	492,430	99.90
2011	463,450	64	241	463,145	99.93
1998-2011	**7,809,003**	**1,227**	**7,064**	**7,800,617**	**99.89**

Author Julia Hipisley-Cox, University of Nottingham
Date: 09.06.2013
Version 1.0

4.3 Hospital Episode Statistics

- **Data Source**: Health & Social Care Information Centre
- **Time Period**: 01 April 1997 to 31 Aug 2012
- **Updates**: annual
- **Description**: includes outpatient referrals (01.03.2003-31.08.2013), admitted patient episodes (01.04.1997-31.08.2012), A&E attendances (01.04.2007-31.08.2012).

4.3.1 Inpatient hospital data 1997-2012

The table below shows in patient admission episodes 1997 to 2012. The percentage with a valid NHS number in 1997/8 was low (45%) but rose quickly to 90% by 2002/3. By 2006, it was above 96.5% - reaching 98.6% by 2012/13.

Table 4: HES Inpatient data: completeness and validity of NHS Number 01April1997 to 31Aug 2012

	Total records	NHS number missing	NHS number invalid	NHS number valid	% NHS number valid
1997/8	11,610,641	4,992,562	1,450,828	5,167,251	44.50
1998/9	12,077,033	3,113,719	9,966	8,953,348	74.14
1999/00	12,723,428	2,489,256	9,273	10,224,899	80.36
2000/1	12,896,485	2,130,697	28,774	10,737,014	83.26
2001/2	12,973,256	1,730,336	2,856	11,240,064	86.64
2002/3	13,442,308	1,277,318	6,040	12,158,950	90.45
2003/4	14,129,373	901,335	5,409	13,222,629	93.58
2004/5	14,546,126	655,899	2,366	13,887,861	95.47
2005/6	15,395,157	640,096	1,467	14,753,594	95.83
2006/7	15,803,643	542,195	1,987	15,259,461	96.56
2007/8	16,456,185	509,395	766	15,946,024	96.90
2008/9	17,434,446	488,560	322	16,945,564	97.20
2009/10	18,126,831	377,136	244	17,749,451	97.92
2010/11	18,727,345	312,385	139	18,414,821	98.33
2011/12	18,889,329	269,484	161	18,619,684	98.57
2012/13	7,901,395	108,640	62	7,792,693	98.62

4.3.2 Outpatient hospital data 2003-2012

For the first year of outpatient data (2003/4), 94% of records had a complete and valid NHS number. For the last 5 years, this has exceeded 98%. By 2012/13 (preliminary data) this had risen to 99%.

Table 5: HES Outpatient data: completeness and validity of NHS Number 01April2003 to 31Aug 2012

year	total records	NHS missing	NHS invalid	NHS valid	NHS valid %
2003/4	51,427,003	3,154,423	5,046	48,267,534	93.86
2004/5	54,420,813	2,373,564	3,891	52,043,358	95.63
2005/6	60,608,403	2,089,818	4,797	58,513,788	96.54
2006/7	63,217,226	1,797,077	116,497	61,303,652	96.97
2007/8	66,649,484	1,493,871	1,175	65,154,438	97.76
2008/9	74,853,493	1,422,921	1,976	73,428,596	98.10
2009/10	84,198,458	1,507,359	850	82,690,249	98.21
2010/11	87,683,207	1,179,740	498	86,502,969	98.65
2011/12	90,956,844	1,047,402	472	89,908,970	98.85
2012/13 (to Aug 2012)	38,262,073	400,584	141	37,861,348	98.95
total	**672,277,004**	**16,466,759**	**135,343**	**655,674,902**	**97.53**

4.3.3 Accident and emergency: 2007-2012

Levels of completeness of the NHS number in A&E data are lower than inpatient data and outpatient data. This is likely to reflect the nature of the population served. However it has risen steadily each year, reaching 94.32% for 2012/13 (preliminary data)

Table 6: HES A&E: completeness and validity of NHS Number 01April 2007 to 31Aug 2012

year	total records	NHS missing	NHS invalid	NHS valid	NHS valid %
2007/8	12,318,051	1,800,928	457	10,516,666	85.38
2008/9	13,794,072	1,537,979	969	12,255,124	88.84
2009/10	15,569,736	1,539,276	513	14,029,947	90.11
2010/11	16,241,015	1,292,160	210	14,948,645	92.04
2011/12	17,619,708	1,138,496	487	16,480,725	93.54
2012/13	7,767,815	440,926	120	7,326,769	94.32

Author Julia Hipisley-Cox, University of Nottingham
Date: 09.06.2013
Version 1.0

4.4 Primary Care (QResearch)

- **Data Source**: We included all 607 practices in England currently contributing to the QResearch database on 1st March 2013.
- **Database version**: We used version 35 of the database (uploaded 6th March 2013).
- **Subjects**: We included all 5,078,704 men and women who were registered on 1st March 2013.
- **Analysis**: 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).
- **Summary**: 99.83% of patients have a complete and valid NHS number.

There were 5,078,704 currently registered patients. Of these, 99.83% had a valid 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.

Table 7: QResearch completeness and validity of NHS Number currently registered patients March 2013

	total patients	patients with valid NHS	% patient valid NHS
all patients	5,078,704	5,070,000	99.83
women	2,563,562	2,559,330	99.83
men	2,515,142	2,510,670	99.82
Type of EMIS System			
EMIS LV	2,407,975	2,405,059	99.88
EMIS Web	2,670,729	2,664,941	99.78
BY Strategic Health Authority			
East Midlands SHA	467,517	467,177	99.93
East of England SHA	444,622	444,326	99.93
London SHA	970,032	965,666	99.55
North East SHA	293,984	293,791	99.93
North West SHA	652,115	651,604	99.92
South Central SHA	474,600	474,086	99.89
South East Coast SHA	388,892	388,446	99.89
South West SHA	608,845	607,926	99.85
West Midlands SHA	439,514	439,012	99.89
Yorkshire and the Humber SHA	338,583	337,966	99.82

5 Summary and conclusion

- Levels of completeness and validity of the NHS number in primary and secondary care routine NHS data are extremely high for nearly all data sources.
- Levels in secondary care are now similar to that recorded in general practices where more than 99.8% of current patients have a valid NHS number recorded.
- The 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 table below summarizes the figures for each data source over all the years available and also for the latest full year of data available.
- ONS Cancer and Deaths Registry data has excellent levels of completeness and validity over 15 and 20 years respectively.
- For hospital episode statistics, the NHS number is valid and complete for 94% of A&E records, and 98.6% of inpatient and 98.9% of outpatients.
- For primary care, 99.8% of current patients have a complete and valid NHS number.

Table 8: Summary of completeness and validity of NHS number recording in six primary and secondary care English NHS data sources

	start	end	total records	total records with valid NHS	% records with valid NHS number
All available years					
A&E	01.04.2007	31.08.2012	75,542,582	68,231,107	90.32
Out patients	01.04.2003	31.08.2012	672,277,004	655,674,902	97.53
In patients	01.04.1997	31.08.2012	233,132,981	655,674,902	97.53
Cancer registry	01.01.1990	31.12.2010	6738358	6705806	99.52
Death registry	01.01.1997	31.12.2011	7,809,003	7,800,617	99.89
Last complete year					
A&E	01.04.2011	31.03.2012	17,619,708	7,326,769	94.32
Out patients	01.04.2011	31.03.2012	90,956,844	89,908,970	98.85
In patients	01.04.2011	31.03.2012	18,889,329	18,619,684	98.57
Cancer registry	01.01.2010	31.12.2010	417,389	416,172	99.71
Death registry	01.01.2011	31.12.2011	463,450	463,145	99.93
Primary Care (QResearch)	01.03.2013	01.03.2013	5,078,704	5,070,000	99.83

- This approach has significant advantages over existing approaches to data linkage where data is extracted from source systems 'in the clear' since it better protects patient confidentiality.
- The OpenP software has been implemented within EMIS systems (55% of GP practices) and TPP practices (around 20% of practices) and is used as a standard approach.
- The OpenP software has also been implemented by two major data controllers of NHS secondary care data (The Health and Social Care Information Centre and the Office of National Statistics).
- This project has demonstrated the utility and scalability of applying pseudonymisation at source to the NHS number recorded in NHS clinical data for purposes of data linkage between primary and secondary care data.
- The OpenP software can be used to undertake similar checks of NHS number validity by any organisation within the source system instead of disclosure of identifiable data. This can be done as a data quality check prior to extraction to determine the utility of the source data for data linkage. Data quality criteria can be set depending on the purpose of the project.
- This evidence supports the utility of the NHS number as a unique and reliable identifier within primary and secondary care records which can be pseudonymised at source and used for multi-source data linkage studies.
- This approach, which is scalable and cost-effective, is intended to promote the public benefits of data sharing whilst protect patient confidentiality.

5.1.1 Acknowledgments

Office of National Statistics, Health and Social Care Information Centre, EMIS practices, EMIS and University of Nottingham for their contribution and expertise in supplying data for this project ad in implementing the OpenP software to enable it.

5.1.2 Funding

This project was funded by QResearch.

5.1.3 Competing Interests

JHC is professor of clinical epidemiology at the University of Nottingham and co-director of QResearch® – a not-for-profit organisation which is a joint partnership between the University of Nottingham and EMIS (leading commercial supplier of IT for 60% of general practices in the UK). JHC is also director of ClinRisk Ltd which produces open and closed source software to ensure the reliable and updatable implementation of clinical risk algorithms within clinical computer systems to help improve patient care. Views presented in this article are those of the author not of any related organisation.

Author Julia Hipisley-Cox, University of Nottingham
Date: 09.06.2013
Version 1.0