Health Data in Practice lecture series

Data linkage: methods and quality assurance

Carol Dezateux

29th September 2020

Institute of Population Health Sciences

wellcome

Queen Mary
University of London
Barts and The London

Learning Objectives

At the end of this lecture you should be able to understand

- The applications and value of linked datasets
- Methods used to link data sets
- The NHS number, its role in data linkage and the use of pseudonymisation at source
- Implications of error in data linkage
- Methods used to quality assure data linkage

What is data linkage?

- Data linkage is an important tool in generating rich information to answer important questions about the causes and outcomes of ill health and its wider determinants across the life course
- Linkage can be

  within datasets

  - to identify multiple records for single individuals (one to many) eg hospital admissions
  - to link multiple individuals by some attribute eg mothers to babies, households (one to many, or many to one respectively)

  between datasets

  - to link different electronic health records eg intensive care datasets to hospital admissions datasets
  - to link health data to vital registration data eg birth or death registrations
  - to link health data to disease registries eg cancer or tissue banks
  - to link health data to ‘administrative data’ eg education data
  - to link health data to research studies eg clinical trials or observational studies following children from birth (birth cohort studies)

Why is data linkage useful?

- Efficient: make use of data collected once but which can be used many times for many purposes
- Enables larger scale studies to answer questions on rarer conditions or rarer demographic attributes (eg homeless, uncommon ethnic groups)
- Can provide unbiased estimates of risk factors (eg abortion and subsequent breast cancer), disease frequency or outcomes
- Can be used to follow up outcomes of people in clinical trials for a long period with fewer drop outs and with less respondent burden

Collaborative Group on Hormonal Factors in Breast Cancer*

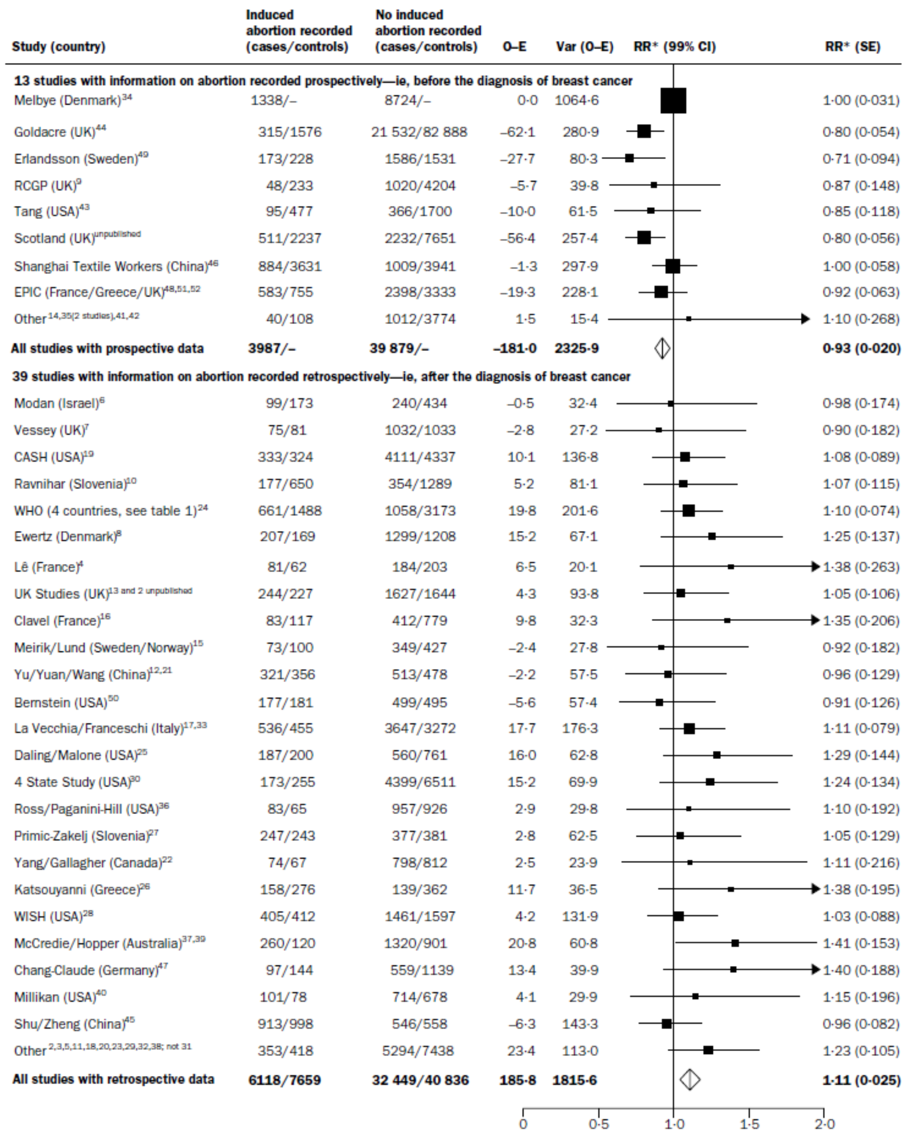
- Collaborative reanalysis of data from 53 epidemiological studies, including 83 000 women with breast cancer from 16 countries

- Studies using prospective information on abortion recorded before the diagnosis of breast cancer were considered separately

- Pregnancies that end as a spontaneous or induced abortion do not increase a woman's risk of developing breast cancer.

- Collectively, studies of breast cancer with retrospective recording of induced abortion yielded misleading results, possibly because women who had developed breast cancer were, on average, more likely than other women to disclose previous induced abortions.

Lancet 2004; 363: 1007–16



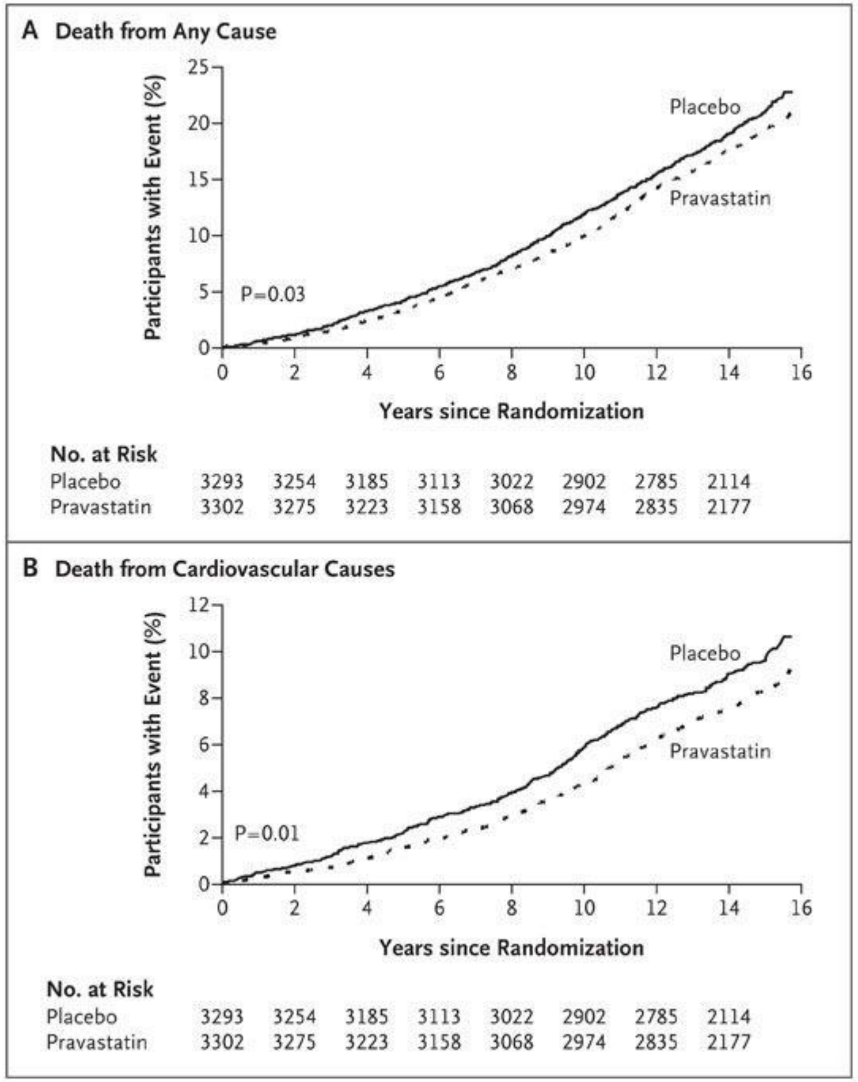
Study (country)	Induced abortion recorded (cases/controls)	No induced abortion recorded (cases/controls)	O–E	Var (O–E)	RR* (99% CI)	RR* (SE)
13 studies with information on abortion recorded prospectively—ie, before the diagnosis of breast cancer						
Melbye (Denmark)[34]	1338/–	8724/–	0·0	1064·6		1·00 (0·031)
Goldacre (UK)[44]	315/1576	21 532/82 888	–62·1	280·9		0·80 (0·054)
Erlandsson (Sweden)[49]	173/228	1586/1531	–27·7	80·3		0·71 (0·094)
RCGP (UK)[9]	48/233	1020/4204	–5·7	39·8		0·87 (0·148)
Tang (USA)[43]	95/477	366/1700	–10·0	61·5		0·85 (0·118)
Scotland (UK)[unpublished]	511/2237	2232/7651	–56·4	257·4		0·80 (0·056)
Shanghai Textile Workers (China)[46]	884/3631	1009/3941	–1·3	297·9		1·00 (0·058)
EPIC (France/Greece/UK)[48,51,52]	583/755	2398/3333	–19·3	228·1		0·92 (0·063)
Other[14,35(2 studies),41,42]	40/108	1012/3774	1·5	15·4		1·10 (0·268)
All studies with prospective data	**3987/–**	**39 879/–**	**–181·0**	**2325·9**		**0·93 (0·020)**
39 studies with information on abortion recorded retrospectively—ie, after the diagnosis of breast cancer						
Modan (Israel)[6]	99/173	240/434	–0·5	32·4		0·98 (0·174)
Vessey (UK)[7]	75/81	1032/1033	–2·8	27·2		0·90 (0·182)
CASH (USA)[19]	333/324	4111/4337	10·1	136·8		1·08 (0·089)
Ravnihar (Slovenia)[10]	177/650	354/1289	5·2	81·1		1·07 (0·115)
WHO (4 countries, see table 1)[24]	661/1488	1058/3173	19·8	201·6		1·10 (0·074)
Ewertz (Denmark)[8]	207/169	1299/1208	15·2	67·1		1·25 (0·137)
Lê (France)[4]	81/62	184/203	6·5	20·1		1·38 (0·263)
UK Studies (UK)[13 and 2 unpublished]	244/227	1627/1644	4·3	93·8		1·05 (0·106)
Clavel (France)[16]	83/117	412/779	9·8	32·3		1·35 (0·206)
Meirik/Lund (Sweden/Norway)[15]	73/100	349/427	–2·4	27·8		0·92 (0·182)
Yu/Yuan/Wang (China)[12,21]	321/356	513/478	–2·2	57·5		0·96 (0·129)
Bernstein (USA)[50]	177/181	499/495	–5·6	57·4		0·91 (0·126)
La Vecchia/Franceschi (Italy)[17,33]	536/455	3647/3272	17·7	176·3		1·11 (0·079)
Daling/Malone (USA)[25]	187/200	560/761	16·0	62·8		1·29 (0·144)
4 State Study (USA)[30]	173/255	4399/6511	15·2	69·9		1·24 (0·134)
Ross/Paganini-Hill (USA)[36]	83/65	957/926	2·9	29·8		1·10 (0·192)
Primic-Zakelj (Slovenia)[27]	247/243	377/381	2·8	62·5		1·05 (0·129)
Yang/Gallagher (Canada)[22]	74/67	798/812	2·5	23·9		1·11 (0·216)
Katsouyanni (Greece)[26]	158/276	139/362	11·7	36·5		1·38 (0·195)
WISH (USA)[28]	405/412	1461/1597	4·2	131·9		1·03 (0·088)
McCredie/Hopper (Australia)[37,39]	260/120	1320/901	20·8	60·8		1·41 (0·153)
Chang-Claude (Germany)[47]	97/144	559/1139	13·4	39·9		1·40 (0·188)
Millikan (USA)[40]	101/78	714/678	4·1	29·9		1·15 (0·196)
Shu/Zheng (China)[45]	913/998	546/558	–6·3	143·3		0·96 (0·082)
Other[2,3,5,11,18,20,23,29,32,38; not 31]	353/418	5294/7438	23·4	113·0		1·23 (0·105)
All studies with retrospective data	**6118/7659**	**32 449/40 836**	**185·8**	**1815·6**		**1·11 (0·025)**

The West of Scotland Coronary Prevention Study

- In this trial, all deaths, hospitalisations and deaths due to coronary events and stroke, and incident cancers and deaths from cancer were tracked with the use of the Scottish record-linkage system

- 5 years of treatment with pravastatin was associated with a significant reduction in coronary events for a subsequent 10 years in men with hypercholesterolemia who did not have a history of myocardial infarction

Ford et al N Engl J Med 2007; 357:1477-1486 DOI: 10.1056/NEJMoa065994



What methods are used to link data?

- There are two main methods used to link data
  - Deterministic or rule-based methods
  - Probabilistic or algorithm-based methods

- Both methods depend on access to identifiers which may be person-based, provider-based or episode-based to identify records which belong to an individual.

- Access to person-based identifiers may be restricted to protect privacy and avoid disclosure of information about patients limiting methods that can be used for linkage in routine health care

- Linkage is never 100% and so we need to quantify the uncertainty or error in linkage and understand its implications

- These will depend on the purposes for which we will be using the linked data

Deterministic methods

- These take a number of identifiers and determine a match based on a combination of these and whether this information is present, is an exact or partial match or is missing
- This method can be used to link many health datasets based on NHS number alone
- However if this is missing, then other identifiers can be used
- Deterministic methods are prone to missed-matches, as any recording errors or missing values may prevent a set of identifiers agreeing
- False-match rates are low, as records belonging to different individuals are unlikely to agree on a complete set of identifiers by chance
- Here is an example

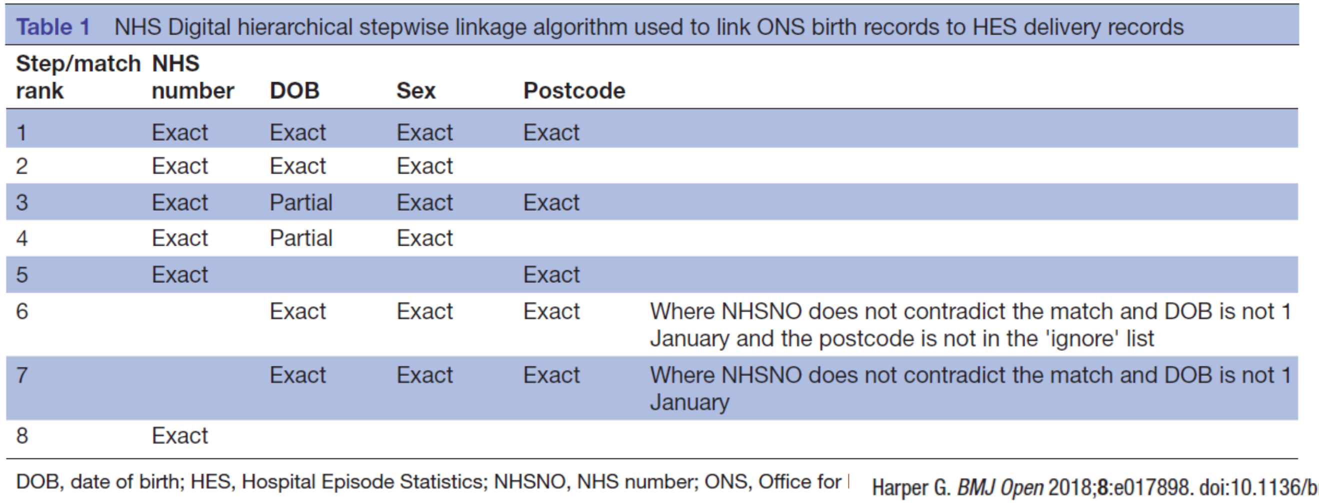


Table 1 NHS Digital hierarchical stepwise linkage algorithm used to link ONS birth records to HES delivery records

Step/match rank	NHS number	DOB	Sex	Postcode	
1	Exact	Exact	Exact	Exact	
2	Exact	Exact	Exact		
3	Exact	Partial	Exact	Exact	
4	Exact	Partial	Exact		
5	Exact			Exact	
6		Exact	Exact	Exact	Where NHSNO does not contradict the match and DOB is not 1 January and the postcode is not in the 'ignore' list
7		Exact	Exact	Exact	Where NHSNO does not contradict the match and DOB is not 1 January
8	Exact				

DOB, date of birth; HES, Hospital Episode Statistics; NHSNO, NHS number; ONS, Office for l

Harper G. *BMJ Open* 2018;**8**:e017898. doi:10.1136/bmjopen-2017-017898

Probabilistic methods

- These assign a weight or a score based on the likelihood of a match and effectively ranks all possible deterministic rules

- A training dataset is usually needed, ideally a gold standard dataset with good quality or complete identifiers

- A number of different statistical approaches are available including the Fellegi-Sunter algorithm based on conditional probabilities, and, more recently, machine learning approaches

- Can reframe linkage error as a missing data problem and impute missing or partially observed data to link and assign match weights

- However these are complex to use in routine practice

Questions???

Unique identifiers in health care: the NHS number

- NHS Numbers are the only unique, national, patient identifier within the UK's health and social care system

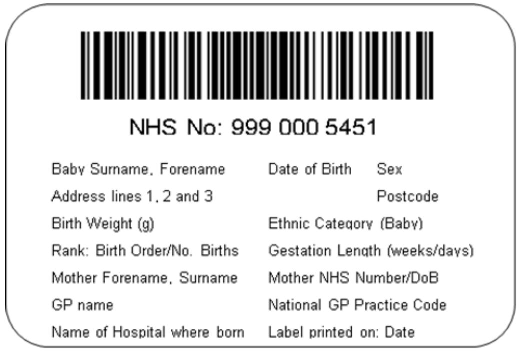
- There are three separate systems in the NHS according to geography/jurisdiction:
  - England, Wales and Isle of Man: NHS number
  - Scotland: Community Health Index
  - Northern Ireland: Health & Care Number

- These numbers provide an important means for linkage

Boyd A, Thomas R, Cornish R, Macleod J. (2018). NHS Number and the systems used to manage them: an overview for research users. Bristol, UK: University of Bristol.

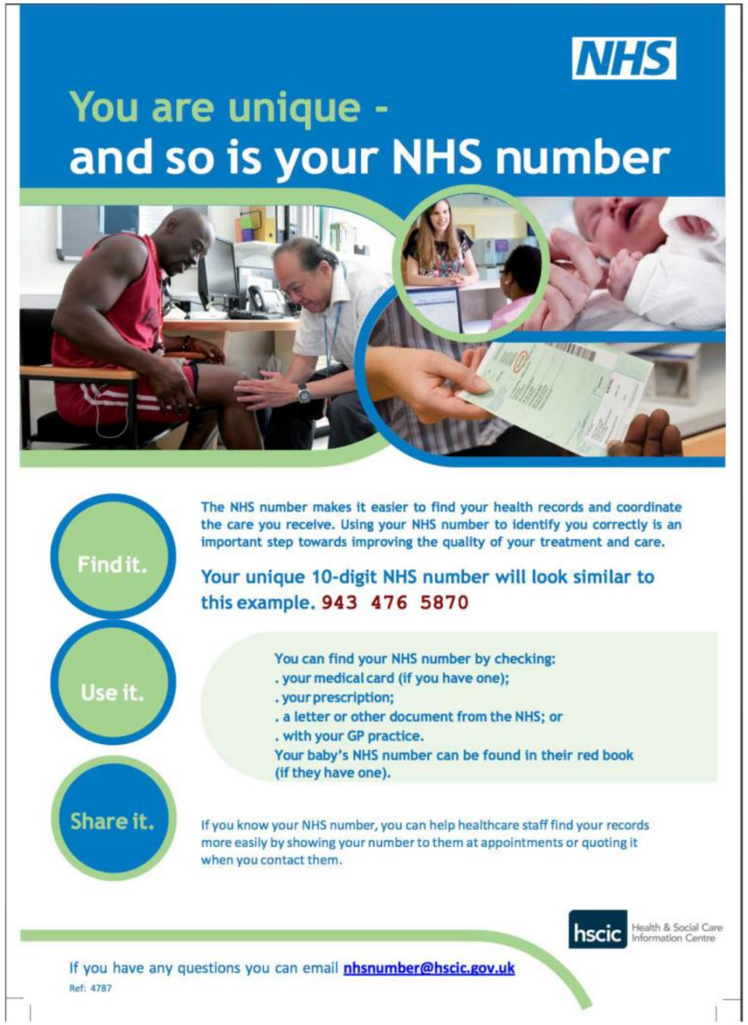
The NHS number in England, Wales and Isle of Man

- A unique 10-digit number displayed as ‘3 3 4’ format eg 123 456 7890
- The first 9 digits are the identifier and the tenth is a check digit used to confirm the number is valid using a Checksum algorithm
- Mandated in all NHS software systems
- The Patient Demographic Service is the NHS service that holds the central register of all allocated NHS numbers
- Since 2002 all babies are allocated a NHS number at birth
- Individuals born outside of these countries or born before 2002 AND not registered with NHS are allocated a number as they join the NHS
- There is a NHS barcode used to label newborn screening specimens or to identify babies by wrist bands



NHS No: 999 000 5451

Baby Surname, Forename	Date of Birth	Sex
Address lines 1, 2 and 3		Postcode
Birth Weight (g)	Ethnic Category (Baby)	
Rank: Birth Order/No. Births	Gestation Length (weeks/days)	
Mother Forename, Surname	Mother NHS Number/DoB	
GP name	National GP Practice Code	
Name of Hospital where born	Label printed on: Date	



NHS

You are unique -
and so is your NHS number

Find it.
Use it.
Share it.

The NHS number makes it easier to find your health records and coordinate the care you receive. Using your NHS number to identify you correctly is an important step towards improving the quality of your treatment and care.

Your unique 10-digit NHS number will look similar to this example. 943 476 5870

You can find your NHS number by checking:
. your medical card (if you have one);
. your prescription;
. a letter or other document from the NHS; or
. with your GP practice.
Your baby’s NHS number can be found in their red book (if they have one).

If you know your NHS number, you can help healthcare staff find your records more easily by showing your number to them at appointments or quoting it when you contact them.

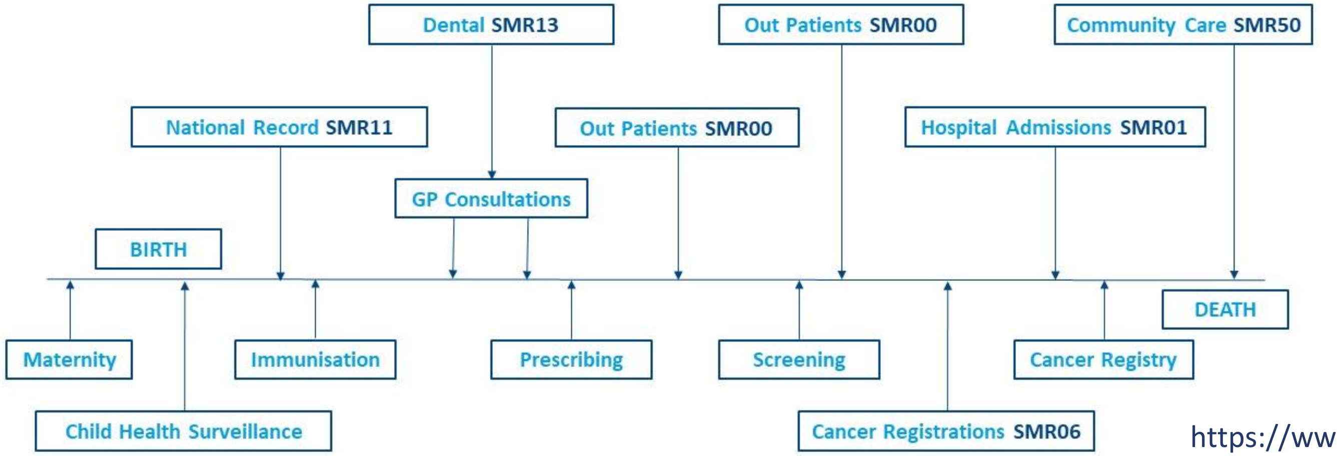
hscic Health & Social Care Information Centre

If you have any questions you can email nhsnumber@hscic.gov.uk

Ref: 4787

The Scottish Community Health Index

- A unique 10-digit numeric code: the first six digits are the date of birth in DDMMYY format; the next 2 are random digits; the 9th is an odd number for males and an even number for females; the 10th is a check digit
- The use of the CHI number is mandatory
- A CHI number is allocated to all NHS Scotland patients who are resident in Scotland, are non-Scottish patients or temporary visitors
- All babies are allocated a CHI number at birth as part of the statutory birth notification
- CHI is used to link a wide range of health data in Scotland

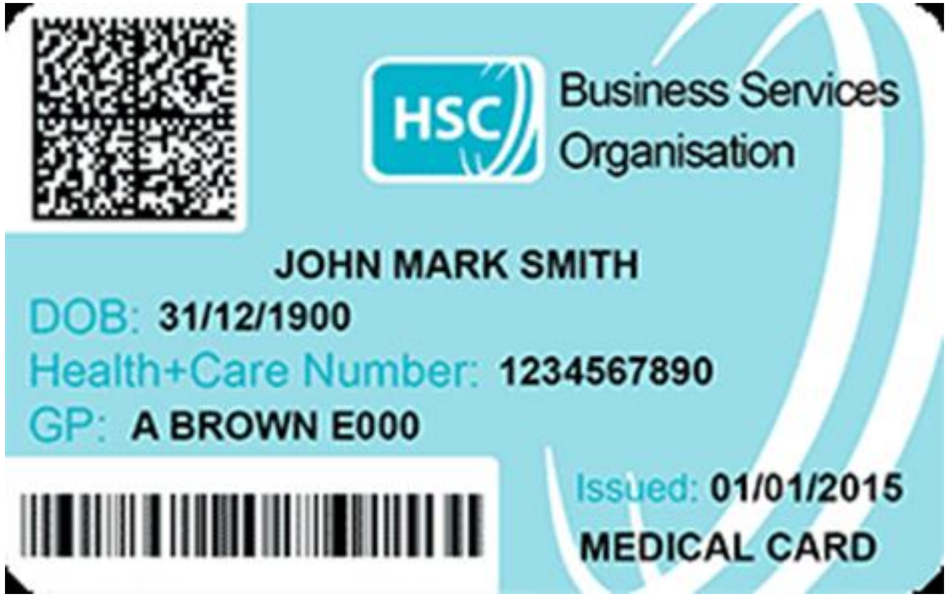


https://www.nhsresearchscotland.org.uk/research-in-scotland/data/sub-page-4

Northern Ireland: Health & Care Number

- The H&C number uses the same 10-digit numerical value as the English NHS Number

- It is presented using the same ‘3 3 4’ format, with the tenth digit being a modulus 11 check-digit
- The H&C Numbers are distinct from NHS Numbers as they are drawn from a reserved range (from 320 000 001 to 399 999 999 plus check digit).
- The H&C Number is being used as the master patient identifier within the Northern Ireland ‘Electronic Care Record’ system



HSC Business Services Organisation

JOHN MARK SMITH

DOB: 31/12/1900

Health+Care Number: 1234567890

GP: A BROWN E000

Issued: 01/01/2015

MEDICAL CARD

How complete is the NHS number in health care datasets?

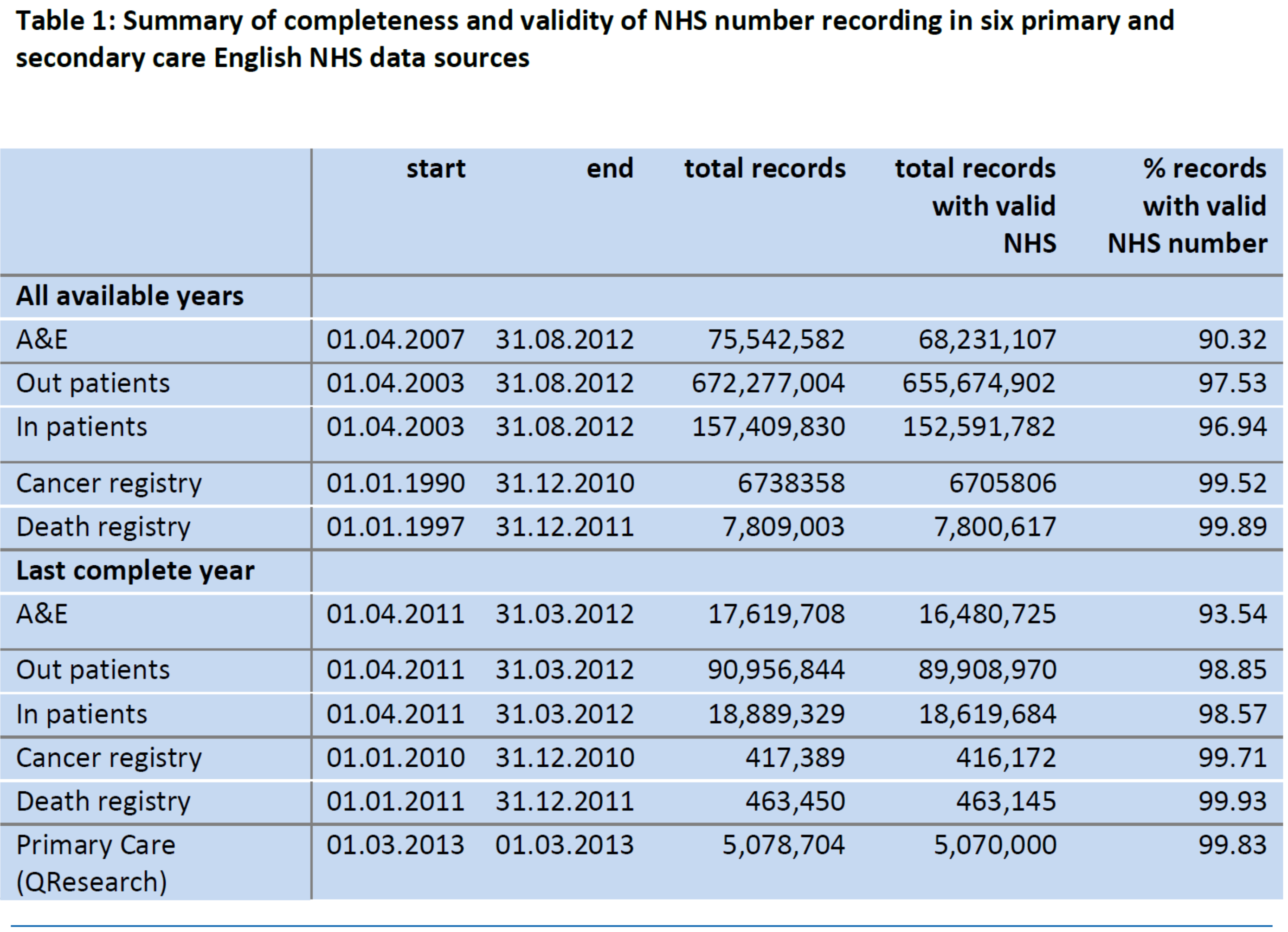
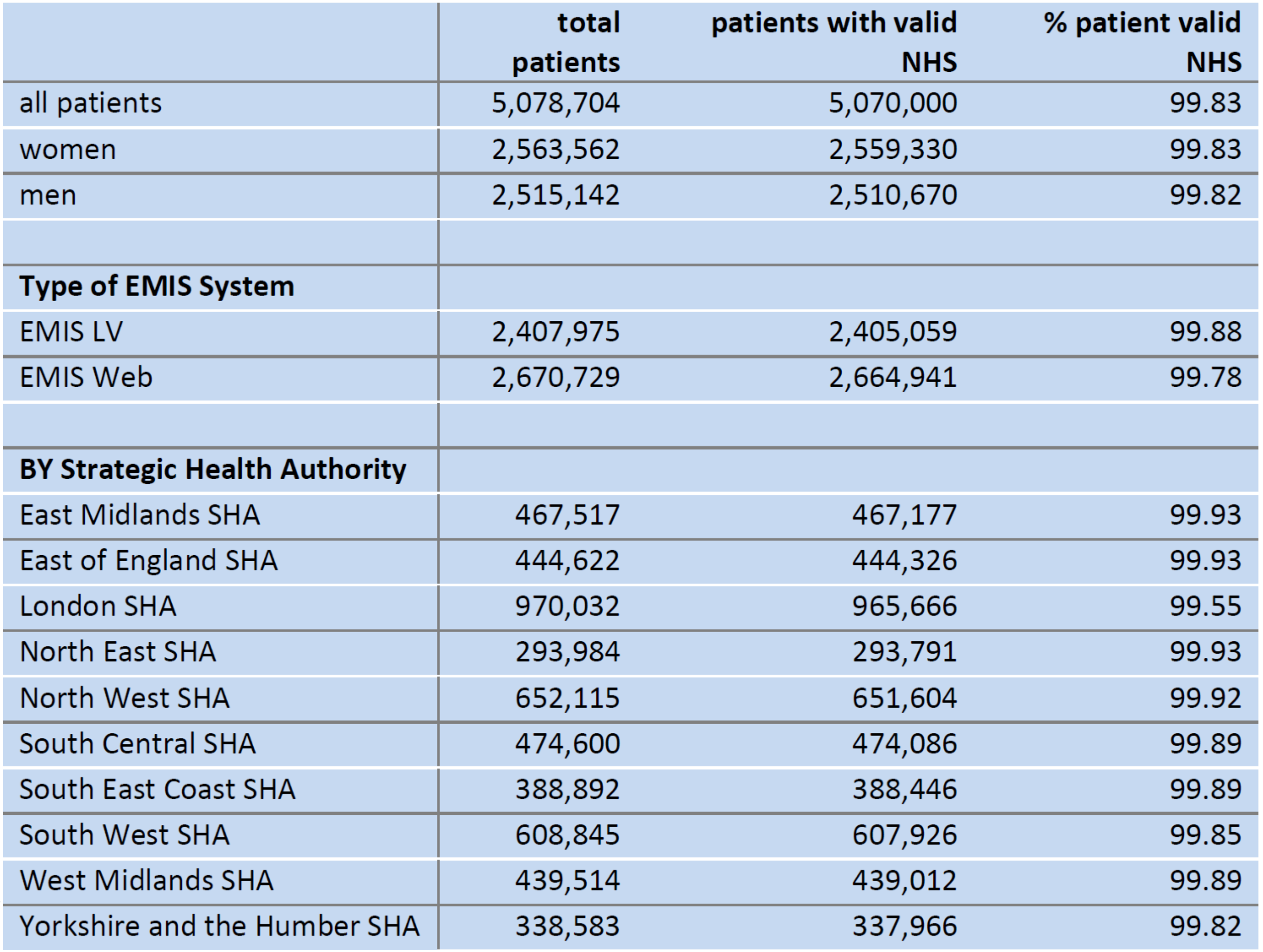


Table 1: Summary of completeness and validity of NHS number recording in six primary and secondary care English NHS data sources

	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	01.04.2003	31.08.2012	157,409,830	152,591,782	96.94
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	16,480,725	93.54
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

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

How complete is the NHS number in health care datasets?



	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

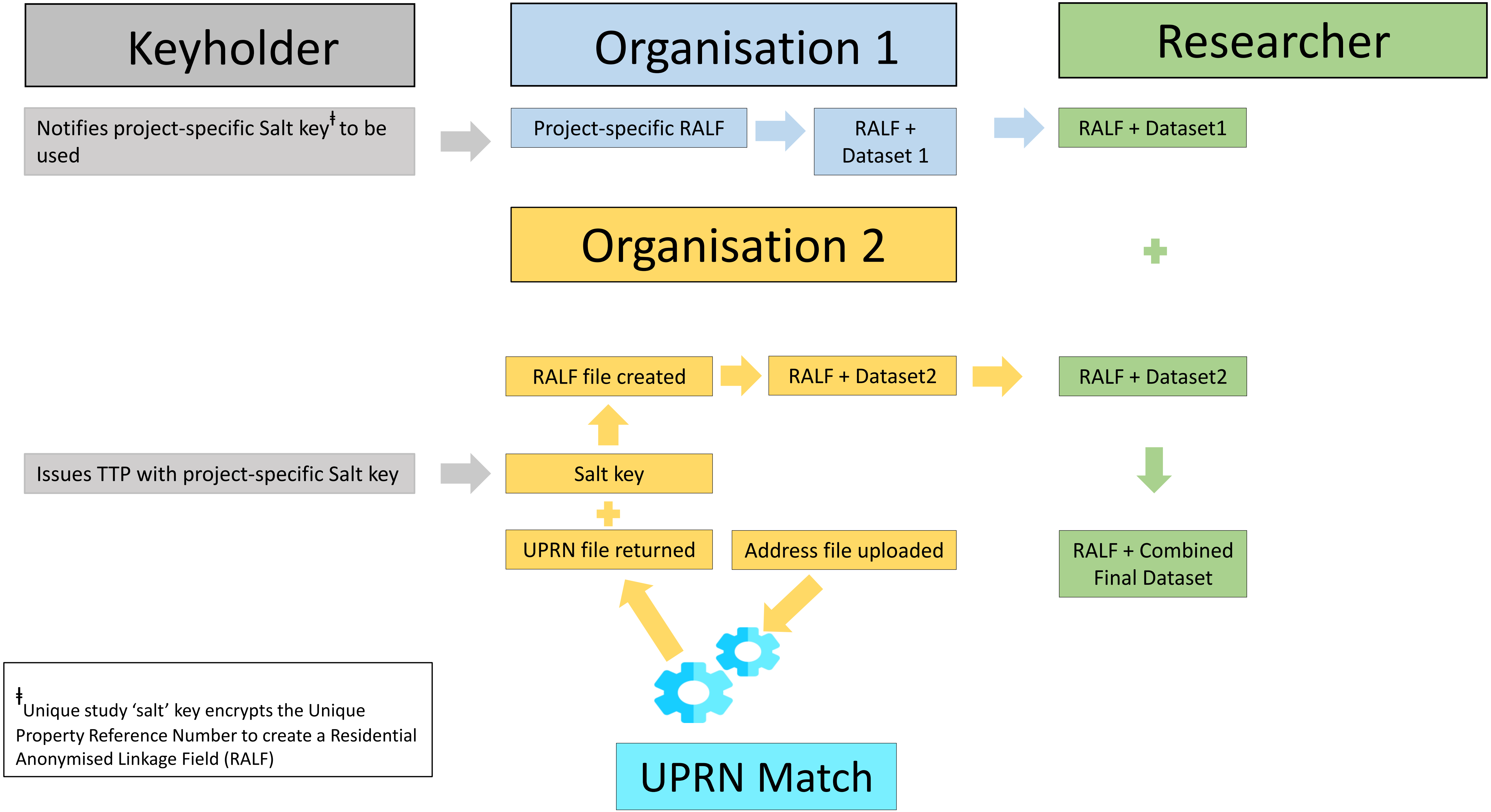
Hippisley Cox 2013 Validity and completeness of the NHS Number in primary and secondary care data in England 1991-2013

Open Pseudonymiser

- Hippisley Cox and colleagues at the University Of Nottingham created an Open Source standalone windows desktop application called OpenPseudonymiser www.openpseudonymiser.org
- Software released as open source for use within NHS clinical computer systems.
- Allows users to pseudonymise datasets by creating a digest of one or more columns of a CSV file
- A digest is the long string created from the input columns in the database and is the key to the pseudonymised data.
- For example, using the column NHS Number of "123456789" would yield the following digest: 15e2b0d3c33891ebb0f1ef609ec419420c20e320ce94c65fbc8c3312448eb225
- Salt is an extra string of characters appended to the data that is being pseudonymised. This allows data to be shared for a specific project with no risk of the data being cross referenced by another project that uses different salt.

Open Pseudonymiser

- Pseudonymisation at source: takes NHS number and creates a project specific encrypted ‘salt code’
- Applies a one-way hashing algorithm (SHA2-256)
- Can be applied twice - before leaving clinical system and on receipt by next organisation
- Identical software can be applied to second dataset using project-specific salt code allowing two or more pseudonymised datasets to be linked
- Can’t be reverse engineered
- Avoids need for storage of look up tables as key can be stored in clinical systems
- As NHS numbers are widely used it is a powerful privacy protecting linkage field
- OpenP is the system used in the Discovery Programme to provide data access for approved third party purposes
- In Discovery it is being used to encrypt an address identifier as well as NHS numbers



Keyholder

Organisation 1

Researcher

Notifies project-specific Salt key[ǂ] to be used

Project-specific RALF

RALF + Dataset 1

RALF + Dataset1

Organisation 2

RALF file created

RALF + Dataset2

RALF + Dataset2

Issues TTP with project-specific Salt key

Salt key

UPRN file returned

Address file uploaded

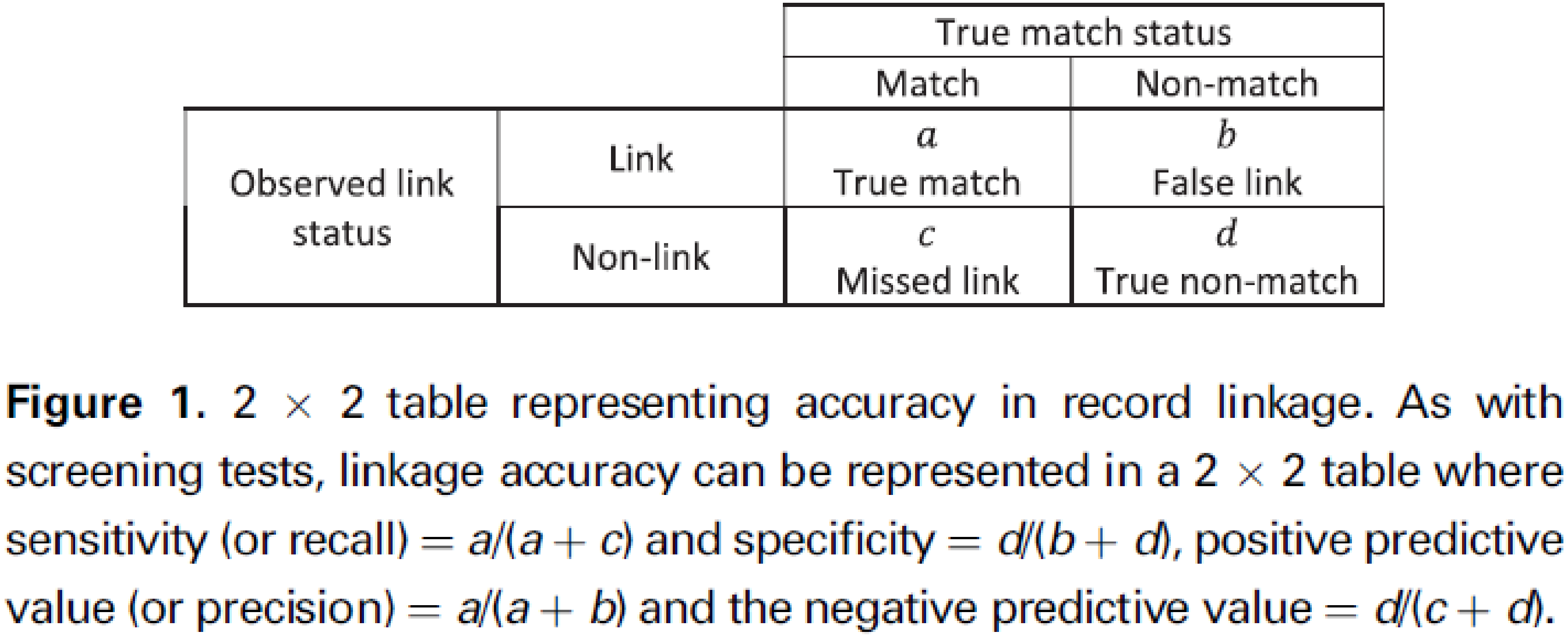
RALF + Combined Final Dataset

UPRN Match

[ǂ] Unique study 'salt' key encrypts the Unique Property Reference Number to create a Residential Anonymised Linkage Field (RALF)

Questions???

How do we measure linkage error or accuracy?



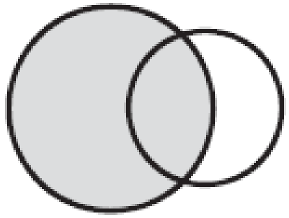
		True match status	
		Match	Non-match
Observed link status	Link	a True match	b False link
	Non-link	c Missed link	d True non-match

Figure 1. 2 × 2 table representing accuracy in record linkage. As with screening tests, linkage accuracy can be represented in a 2 × 2 table where sensitivity (or recall) $= a/(a + c)$ and specificity $= d/(b + d)$, positive predictive value (or precision) $= a/(a + b)$ and the negative predictive value $= d/(c + d)$.

Reproduced from Doidge et al International Journal of Epidemiology, 2019, 2050–2060 doi: 10.1093/ije/dyz203

Common Linkage Structures

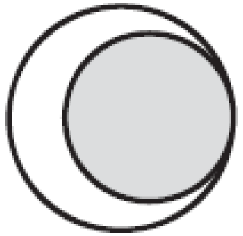
“Master”



- Purpose: to define outcome – for example, all patients registered with general practitioners in east London linked to cancer registry; used to indicate whether the patient has cancer or not. Analyses of shaded area will allow comparison of patients with and without cancer.
- Missed matches may lead some patients who have cancer not being correctly identified and allocated to cancer free group
- False matches may lead some patients who are cancer free being wrongly included in cancer group
- This misclassification weakens or attenuates the associations you may wish to study

Common Linkage Structures

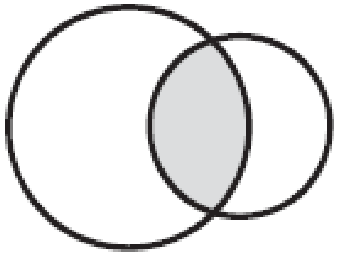
"Nested"



- Purpose: to add further information to the study population of interest define outcome – for example, all Millennium Cohort Study participants in Wales linked to the Welsh NHS immunisation records: this additional information enabled immunisation timeliness to be examined within the cohort study (shaded area) to examine predictors of timely immunisation
- Missed match: no information on those without a linked record
- False match: measurement error in the immunisation variables obtained through linkage

Common Linkage Structures

“Intersection”



- Purpose: to create a study population – for example, the first dataset comprises school attainment records for whole of England and second dataset a research study. Only those with information from both sources are included in the analyses. The overlap defines the study population
- Missed matches may lead to ‘selection’ bias ie the characteristics of those who have linked records may differ systematically from those who do not have linked records
- False match: measurement error in the school attainment variables obtained through linkage

Methods used to assure linkage quality

Gold standard or reference dataset

- true match state is known and can be used to test linkage algorithms and calculate error rate

Post linkage data validation

- Estimate minimum false match rate by identifying implausible scenarios eg pregnant men, resurrections – death date before birth date in data

Sensitivity analyses

- Vary the algorithm or threshold for a match and comparing effect on linkage rates

Comparing linked and unlinked data

- To see whether there are differences in characteristics of individuals by time place and person according to whether linked; compare with external data to assess whether estimates are consistent with other data

Questions???

Summary

- Linked datasets are widely used in health data research
- Deterministic methods are most commonly used to link data based on NHS number however for linkage to other datasets which lack a common unique identifier probabilistic methods are available although can be complex to use
- The NHS number is a unique identifier widely used in in data linkage enabling pseudonymisation at source
- It is important to assess errors in data linkage as people may be missing based on risk factors, demographic factors or outcomes relevant to the research question
- Methods used to quality assure data linkage may help refine the linkage process
- Errors or biases in linkage should be taken into account when analysing or interpreting studies using linked data

Tutorial

We will work through your case study using the NASSS framework tool

You will be assigned an exercise to bring for discussion at next week's tutorial based on today's lecture on linkage

This slide set will be made available to you on QMPlus together with a reading list