

Record linkage using probability matching and the challenges of limited identifying information

Catherine Storey

Linking records

- **Need to be confident records belong to the same person**
- **Ideally need:**
 - Surname and soundex code
 - First name
 - Sex
 - Date of birth
 - Post code
 - Others eg unique CHI number, case reference number

Linking records – Examples (1)

Surname	Soundex	Forename	Date of birth	Post code	CHI number
Duck	D200	Donald	1 January 1955	DL1 2BY	0101551234
Dukk	D200	Donald	1 January 1955	DL2 3SL	0101551234
✘	✔	✔	✔	✘	✔

Surname	Soundex	Forename	Date of birth	Post code	CHI number
Swan	S500	Daisy	24 December 1959	AZ1 2BY	2412591234
Duck	D200	Daisy	24 December 1959	DL2 3SL	2412591234
✘	✘	✔	✔	✘	✔

Linking Records – Examples (2)

- **Both the above would be considered good links**
- **Change of details such as postcode, surname taken into account**
- **Misspelling of surname can be compensated by use of soundex**
- **Unusual surnames given more weight**
- **Constants such as date of birth can be subject to error – allowances made**

A more challenging linkage

- **Request to link Scottish Morbidity Record (SMR) to the Scottish Drugs Misuse Database (SDMD)**
- **SDMD data highly confidential – limited person identifying information:**
 - No surname available – 1st and 4th characters only; as a result, no soundex
 - Forename reduced to initial only, no second initial
 - Postcode sector only
 - No ‘extras’ such as maiden name

Probability matching: making the best use of available data

- Full date of birth and sex present
- Part postcode still useful
- Weight combinations of surname 1st and 4th letter
- Add weight for patients with a history of drugs misuse in SMR – psychiatric records added for this purpose

Managing without a surname

- **Use similar principle to soundex weighting**
- **Assign binit weights according to frequency of character combination in Scottish population – Community Health Index used**
- **Program written to read same character combinations from SMR**
- **Disadvantages**
 - does not compensate for misspelling as well as soundex
 - Difficult to judge at checking stage

Calculating binit weights

Character combination	Frequency of combination in population	Percentage of combination in population	Odds ratio - estimated agreement % (97) divided by percentage in population	Binit weight = log to base 2 of odds ratio
A	5,201	0.096	1012.427289	9.98
AA	25,887	0.477	203.4084417	7.67
AB	862	0.016	6108.624513	12.58
AC	2,998	0.055	1756.382365	10.78
AD	3,280	0.060	1605.37632	10.65
AE	50,103	0.923	105.0961885	6.72
AF	788	0.015	6682.277069	12.71
AG	333	0.006	15812.71571	13.95
AH	8,423	0.155	625.1495109	9.29
AI	12,033	0.222	437.5994623	8.77
AJ	178	0.003	29582.21534	14.85
AK	8,199	0.151	642.2288486	9.33
AL	3,133	0.058	1680.700393	10.71

Incorporating patient history

- Intention to increase probability of good match, compensate for lack of identifiers
- Psychiatric records included in linkage to capture drugs misuse
- ICD-9 codes 292, 304, 305
- ICD-10 codes F10-F19, T40
- Bonus of 3 points given to add weight to patients with history of drugs misuse

Initial results – checking the paired records

- **Vital stage in setting a threshold for the linkage**
- **Judgement more difficult in absence of full surnames and postcodes**
- **More attention given to potential misspelling of surname**
- **Ensured that ‘drug’ bonus was not falsely enhancing score by being too high, or eliminating good matches with no history**

Checking the paired records – examples (1)

Example 1: Strong all-round score without 'drug' bonus

Type	Surname	Forename	Date of birth	Sex	Post code	Drug
SMR1	POTTER	J	19750101	M	G1 8BZ	
SDMD	PT	J	19750101	M	G1 8	No

Scores:

Total	Surname	Forename	Date of birth	Sex	Post code	Drug
34.90	7.71	2.28	14.76	1.00	9.15	0.00

Example 2: Lower score enhanced by 'drug' bonus

Type	Surname	Forename	Date of birth	Sex	Post code	Drug
SMR1	PIERCE	M	19701231	M	EH1 9RU	
SDMD	PR	M	19701231	M	EH	Yes

Scores:

Total	Surname	Forename	Date of birth	Sex	Post code	Drug
34.81	8.37	4.27	14.76	1.00	3.41	3.00

Checking the paired records – examples (2)

Example 3: Possible misspelling of surname

Type	Surname	Forename	Date of birth	Sex	Post code	Drug
SMR1	MACDONALD	G	19501231	M	G6 6PS	
SDMD	MO	G	19501231	M	G6 6	Yes

Scores:

Total	Surname	Forename	Date of birth	Sex	Post code	Drug
26.80	-5.00	3.89	14.76	1.00	9.15	3.00

Example 4: Possible incorrect character in date of birth

Type	Surname	Forename	Date of birth	Sex	Post code	Drug
SMR1	SMITH	D	19800831	M	EH165JY	
SDMD	ST	D	19800331	M	EH16	Yes

Scores:

Total	Surname	Forename	Date of birth	Sex	Post code	Drug
28.79	8.82	3.56	5.20	1.00	7.21	3.00

Setting the threshold

- **Around 200 sampled pairs checked at scores from 24 to 31**
- **Threshold set at score at which it was more likely that pairs matched than did not match – 26**
- **Linkage program re-run with this threshold**

Results of the linkage

- **Of 146,993 SDMD records, 115,287 linked to SMR – 78.4%**
- **Final data set included**
 - 292,516 hospital discharges
 - 39,748 psychiatric discharges
 - 1,450 cancer registrations
 - 3,193 deaths