10. レコードリンケージにおける重複者チェックの効果的方法

1. 緒　言

情報をわれわれが利用できる資源のうちで最も重要なものの1つと見なす考えは，古くから提唱されてきた。一般に，情報という資源は，広い分野に散在しており，人間の健康や疾病に関する情報においても例外ではない。個人の健康に関する情報を，疫学的研究に十分に活用するためには，異なる時間や異なる場所で記録された複数の情報を結合させること（レコードリンケージ）が必要である。しかし，レコードリンケージを実施するにあたっては新しく発生したデータをいかにして患者固有のセグメントに結合させるかという個人識別の問題がある。個人識別を迅速かつ正確に行なうためには，ファイル中に同一個人の重複がないことが必要である。重複者のチェックを行なう1つの方法として，ファイル中のデータ（個人）の任意の組み合わせ（2人1組）についてチェックを行なう方法（以後，総当り法と呼ぶ）が考えられる。しかしながら，登録件数が10万件以上の大容量のファイルにおいて，総当たり法で重複者のチェックを行なうと，高速のコンピュータを使用したとしても処理時間は膨大になり，事実上実施不可能と思われる。

われわれは，大容量のファイルにおいて，実施可能な重複者チェックの方法の開発を進め，登録件数が約10万件のファイルでも数時間のコンピュータ処理で行なえ，しかも重複者の発見率について，総当り法とほとんど差がない方法を開発した。また，この方法の適用範囲を知るために，実際に重複が確認された組み合わせについて，重複の原因およびパターンの分布について分析を行なった。

2. 資　料

われわれは，原爆被爆者診療記録データベース作成業務として，日赤長崎原爆病院の内科入院記録のコンピュータファイル化を進めている。1980年の時点で5,160名の被爆者について，姓名，性別，生年月日および被爆時状況からなる個人情報と入退院日，入院時検査値，診断名などからなる診療情報の登録を完了している。しかしながら，5,160名の登録者の中には同一人物が重複して登録されている可能性がある。重複登録には個人同定に関する情報が全て一致している場合と，何らかの理由によりその一部が異なる場合とがある。本研究では，個人情報のうち姓名，性別，生年月日を個人同定の情報として使用した。

3. 方　法

同一人物であるか否かを判定するためのチェック項目として，姓および名（カナ文字），性別，出生年号，出生年，出生月，出生日の7項目をチェック項目とした。姓および名において，濁点と半濁点によるフリガナの間違いが多いので，濁点と半濁点を姓および名のカナ文字から除き，上記の7項目と患者番号からなるマスターファイルを作成した。このマスターファイルを用いて重複者チェックを行なうために，ソーティングを用いる方法

（以後，分類法と呼ぶ）を開発した。分類法と総当り法を用い同一人物と思われる組み合わせを出力し，この組合せについて入院記録により本籍と既往歴等を調べ，同一人物であるか否かの判定を行なった。

(1) 分類法

種々のソートキーでソーティングを複数回行ない，おのおのの場合の隣接者のみについてチェック項目を比較し，検出基準を満たす組み合わせを出力する。但し，各ソーティングとも性別を第1ソートキーとした。

(2) 総当り法

マスターファイル中から任意の2人を取り出すすべての組み合わせについて，7項目の比較を行ない，一致した項目ごとに定められた点数を与え，その合計をその組み合わせの得点とした。この得点が基準点以上の組み合わせを出力した。各項目の点数は姓と名に各3点，性別，出生年号，出生年，出生月，出生日に各1点，さらに出生年号と出生年月日がすべて一致した場合にのみ上記点数とは別に2点を与えて，全項目が一致した場合の得点を13点とした。

4. 結　果

(1) 発見率

表1は分類法により重複が発見された組み合わせ207組について，実際に行なった10回のソーティングのソートキーと検出基準及び出力組数と重複組数を示したものである。重複者の中には複数回検出される者も含まれていた。表2は総当り法により重複が発見された組み合わせ209組について得点別に出力組数と重複組数を示したものである。分類法の方で重複が発見された組み合わせ数が総当り法より2組少ないのは，性別が誤登録されていた組み合わせについて発見できなかったためである。

(2) 発見効率

作業の効率という意味で，重複組数／出力組数を発見効率と定義する。分類法では，全体をとおして207/901＝23.0（%）であり，総当たり法では，全体で209/912＝22.9(%)で両者間にはほとんど差はなかった。

(3) 処理時間

分類法では1回のソーティングの処理時間は4分以内であり，10回合わせた処理時間は1時間以内であった。総当り法では約7.5時間であった。分類法の処理時間は総当り法の1/7以下であった。

(4) 重複のパターン

重複登録された組み合わせは，チェック項目中のいずれかの項目が不一致になっているので，項目ごとに一致不一致を調べ，そのパターンと度数を表3に示した。項目別では名の不一致が最も多く209組中の107組であった。次いで姓の80組，出生日が11組，出生月が8組，出生年号および出生年が4組，性別が2組であった。不一致の項目数については，1個が183組，2個が28組で重複登録者のうち約1割は7個のチェック項目中2個が同時に不一致となっていた。

(5) 重複登録の原因別分類

重複登録された組み合わせについて，不一致項目が生じた原因を原票の誤り，改姓，改名，読み違い，その他（コーディング，パンチの誤り）に分類して項目ごとにその度数を調査し，その結果を表4に示した。同一漢字が異なる読み方をされたことが最も多く(112組)，次いでコーディングおよびパンチ上の誤まりが（83組）で，両者で全体の9割以上

を占めていた。しかしながら，原票の時点での誤まりが（15組）も含まれていたことは注意すべきことである。

5. 考　察

総当たり法で発見され，分類法で発見できなかった2組の重複登録者は，いずれも性別の誤まりと名の読み違いが重なったものであった。管理された医療データのように情報がいっそう正確なものであれば，分類法でも重複登録が発見されると考えられる。

二つの方法の処理時間の差は，チェックの実行回数の差（すなわち，$_{5160}C_2$ —5159×10＝13310220—51590＝13258630）に依存し，データ数が増えるにつれて，急速に増大する。

重複登録の原因は姓および名の読み違いによるものが全体の半数以上であり，原票にフリガナをつけるよう義務づけることにより，重複登録を大幅に減じることができると考えられる。

6. 結　論

レコードリンケージを行なう上で必要不可欠な重複登録者を発見するための手段として分類法を開発し，総当り法との比較を行なった。分類法は処理時間がはるかに短かく，総当り法で発見できる重複登録者の99%を発見でき，発見効率についても両者間には差はなかった。また，管理された医療データ・ベースがいっそう正確なものであれば，さらに良い結果を得るであろう。

これらの結果と重複の原因の分析を行なうことによって，分類法は有効であり，とくに総当り法が事実上実施不可能となるような大容量ファイルの場合にも，容易に重複者のチェックを行なうことができることを知った。また，複数のファイル間でレコードリンケージを行なう場合に，それらのファイルを1つのファイルにまとめ，分類法により重複者のチェックを行なうことにより，結合すべきデータの組み合わせを容易に発見できると考えられる。

（近藤　久義）

表 1. 分類法の発見効率

	ソートキー			検出基準	出力組数	重複者組数	発見効率(%)
	第二	第三	第四				
1	姓	年号,年,月		姓，性別，年号，年，月が一致	173	113	65.3
2	姓	年号,年	日	姓，性別，年号，年，日が一致	135	110	81.5
3	姓	年号	月，日	姓，性別，年号，月，日が一致	147	108	73.5
4	姓	年,月,日		姓，性別，年，月，日が一致	113	109	96.5
5	名	年号,月,日		名，性別，年号，年，月が一致	140	90	64.3
6	名	年号,年	日	名，性別，年号，年，日が一致	99	85	85.6
7	名	年号	月，日	名，性別，年号，月，日が一致	120	84	70.0
8	名	年,月,日		名，性別，年，月，日が一致	92	86	93.5
9	姓	名		姓，姓，性別が一致	74	33	44.6
10	年号,年,月,日			性別，年号，年，月，日が一致	637	179	28.1
合計					901	207	23.0

注）年号：出生年号，年：出生年，月：出生月，日：出生日

表2. 総当り法の発見効率

得点	誤登録項目の組み合わせ	出力数	重複数	発見効率(%)
6	(1)姓＋名＋性別 (2)姓＋〔年号，年，月，日〕の二つ (3)名＋〔年号，年，月，日〕の二つ (4)性別＋年号＋年＋月＋日 (5)性＋性別＋〔年号，年，月，日〕の一つ 名＋性別＋〔年号，年，月，日〕の一つ	3501	0	0
7	(1)姓＋名 (2)姓＋〔年号，年，月，日〕の一つ (3)名＋〔年号，年，月，日〕の一つ (4)性別＋〔年号，年，月，日〕の三つ (5)年号＋年＋月＋日	698	24	34.4
8	(1)性別＋〔年号，年，月，日〕の二つ (2)〔年号，年，月，日〕の三つ	19	0	0
9	(1)姓＋性別 (2)名＋性別 (3)性別＋〔年号，年，月，日〕の一つ (4)〔年号，年，月，日〕の二つ	11	4	36.5
10	(1)姓 (2)名 (3)〔年号，年，月，日〕の一つ	164	160	97.6
12	(1)性別	0	0	-
13	合項目一致	21	21	100.0
合計	得点7点以上	913	209	22.9

注) 1. 年号：出生年号，年：出生年，月：出生月，日：出生日

2. 得点が13点の組み合わせは，姓または名のフリガナの濁点・半濁点の有無により，異なる人物として登録されたもの。

表3. 誤登録項目パターン

	姓	名	性別	出生時				重複複登録	
				年号	年	月	日	組数	(%)
11	I	I	I	I	I	I	I	21	10.0
2	O	I	I	I	I	I	I	64	30.5
3	I	O	I	I	I	O	I	86	41.0
4	O	O	I	I	I	I	I	11	5.3
5	I	O	O	I	I	I	I	2	1.0
6	I	I	I	O	I	I	I	4	1.9
7	I	I	I	I	O	I	I	1	0.5
8	I	I	I	I	I	I	O	5	2.4
9	O	I	I	I	O	I	I	1	0.5
10	O	I	I	I	I	O	I	2	1.0
11	O	I	I	I	I	I	O	2	1.0
12	I	O	I	I	O	I	I	1	0.5
13	I	O	I	I	I	O	I	5	2.4
14	I	O	I	I	I	I	O	2	1.0
15	I	I	I	I	O	I	O	1	0.5
16	I	I	I	I	I	O	O	1	0.5
合計	(80)	(107)	(2)	(4)	(4)	(8)	(11)	209	100.0

注) O：不一致，I：一　致

合計欄のカッコ内の数字は各項目の不一致に対応する重複組数を示す

表4. 誤登録の原因別の頻度と重複割合

	誤登録項目	原因		重複数	重複割合(%)
1	姓	濁点の有無		21	10.0
2	姓	読み違い		25	12.0
3	姓	改姓姓		9	4.3
4	姓	その他		30	14.4
5	名	読み違い		48	23.0
6	名	その他		38	18.2
7	出生年号	その他		4	1.9
8	出生年	原票		1	0.5
9	出生日	原票		5	2.4
10	出生年，出生日	原票		1	0.5
11	出生月，出生日	原票		1	0.5
12	性，名	読み違い	読み違い	9	5.3
		改姓	読み違い	1	
		改姓	改姓	1	
13	姓，出生年	読み違い，その他		1	0.5
14	姓，出生月	原票，その他		2	1.0
15	姓，出生日	読み違い，原票		2	1.0
16	名，出生年	原票，その他		1	0.5
17	名，出生月	読み違い	その他	2	2.4
		その他	その他	3	
18	名，出生日	読み違い	原票	1	1.0
		原票	原票	1	
19	名，性別	読み違い，その他		2	1.0

注）その他：コーディング，パンチの誤り