千葉衛研報告　第29号　66－70　2005年　　(資料)

おたっしゃ調査における突合作業について

茂野　誠一，柳堀　朗子，須田　和子，一戸　貞人

Record linkage on cohort study

Seiichi SHIGENO, Ryoko YANAGIBORI,
Kazuko SUDA, Sadato ICHINOHE

1．はじめに

健康疫学研究室は，鴨川市で実施中の疫学調査（おたっしゃ調査：前向きおよび後ろ向きコホート研究）のデータ管理を担当している。データ管理には，突合（レコードリンケージ）と呼ばれる作業が含まれる。今回，我々が平成16年度に実施した突合作業について報告する。

2．コホート調査とは

コホート研究では，疾病発症の要因を突き止めるために，特定の集団を追跡し，個人を特定した上で疾病発生情報とその要因となる生活習慣などの複数の情報を収集して解析を行う疫学調査のことである。

3．突合（レコードリンケージ）とは

一般的に，突合（レコードリンケージ）とは，2つあるいはそれ以上の記録（異なった診療録や出生証明書，死亡診断書などの人口動態記録）に含まれる情報を収集し組み合わせることであり，このため，同一個人が1回しか数えられないように個人を識別する手続きを必要とする。

突合は，1940年代後半には人手による個人同定作業が行われ[1]，その後大型計算機が用いられるようになり[2]，現在では，突合用のソフトも存在する[3]。その間に確率論的な手法（probabilistic record linkage）が導入されている。今回我々は，ノートパソコンを使用してデータベースソフトと表計算ソフトを用いて作業を行った。

4．おたっしゃ調査とは

「おたっしゃ調査」は，「高齢になっても寝たきりや痴呆にならないための予防策」を立てる基礎資料とし，脳卒中や心臓病，糖尿病などの病気と生活習慣の関連を解明することを目的として千葉県と鴨川市（旧鴨川市，旧天津小湊町）が共同で実施している。基本健康診査(総合検診)結果が保管されている鴨川市（旧鴨川市，旧天津小湊町）で，平成16年1月時点で40歳以上となる住民を対象とした平成20年度までの5年間追跡を行うコホート研究である。調査協力に同意した住民について個人情報保護対策を厳重に行った上で，以下に述べるデータについて，追跡期間終了後に解析を実施する。対象者23,073人のうち生活習慣アンケート回収数10,470件（回収率46.5%），性年齢の判明している有効回答数は10,127人，検診結果，介護保険の認定状況の調査に協力しコホート研究の対象となる者は，6,511人であった。この調査協力承諾者の生活習慣アンケート結果と総合検診結果，要介護認定状況，疾病発症状況等をと都合した上で解析が行われる。

5．おたっしゃ調査における突合作業

1）作業の方針

平成15年度，16年度，昭和62年度に総合検診を受けた住民から調査協力承諾者を確定するために，総合検診結果の一覧から個人識別情報のみを取り出し，調査協力承諾者リストの個人識別情報と照合し一致ペアを確定することで作業を行うこととした。データ照合と検証の手順は，異種データベース間でのレコード照合に関する研究動向を参照した[1]。作業はエクセル形式で保存されたデータを使用し，マイクロソフトアクセスを用いてデータ抽出を行い，マイクロソフトエクセルで検証を行うこととした。

2）予測された問題点と対応

作業開始前の確認で，三つの問題点が見つかっていた。第一に，検診結果リストと承諾者リストでは記録様式が一部異なっていること（表－1）。第二に，承諾者リストには，氏名に含まれる文字がフォントの関係で入力できない場合○となっていること。第三に昭和62年度の検診結果では，氏名がカタカナ表記であること。

表－1　旧市町の検診結果と調査協力承諾者リストの記録フォーマットの違い

旧市町検診結果	氏名(*1)	生年月日(*2)	住所(*3)	
	千葉　一郎	19581001	仁戸名町666-2	
調査協力承諾者リスト	氏名	生年月日	地区名	番地
	千葉一郎	1958/10/1	仁戸名町	666-2

(*1) 姓と名の間をスペースで区別。(*2) 生年月日を文字列で入力。
(*3) 住所を一体で入力。

千葉県衛生研究所
（2006年1月31日受理）

記録様式の不一致への対応は，それぞれエクセルを用いてデータを加工することとした。

(a) 検診結果の姓と名の間のスペースを除いた(図－1)。

(b) 検診結果の生年月日に"/"付加する(図－2)。

(c) 住所の地区名と番地を分離する(図－3)。

(d) 異体字，変体仮名等の入力できなかった文字の問題対策として氏名の書く文字を分離したリストを作成行い一文字での照合を可能とする(図－4)。

(e) 昭和62年度分の健診結果と照合するため，氏名をカナ表示とした承諾者リストを作成する。

A	B	C	D	E	F
1	NO.	氏名（オリジナル）	氏名（変換後）	変換式	注釈
2	1	千葉　一郎	千葉　一郎	=SUBSTITUTE(B4," ","")	全角スペースを変換
3	2	千葉　一郎	千葉　一郎	=SUBSTITUTE(B4," ","")	半角スペースは変換されないので，再度実施する必要がある(*)

(*) No2は，半角スペースを2個用いて氏名を区切っていたため，返還式=SUBSTITUTE(B4," ","") の「" "」の"と"の間を半角スペースで指定する必要があった。

図－1　エクセルによる氏名の間のスペースの除去

A	B	C	D	E	F	G	H
1	NO.	生年月日	文字列1	文字列2	文字列3	文字列結合結果	文字列→数値
2	文字列	19250101	1925	01	01	1925/01/01	1925/1/1
3			=LEFTB (B2,4)	=MIDB (B2,5,2)	=RIGHTB (B2,2)	=C2&"/"&D2&"/ "&E2	=VALUE(G2)

G3のセルの文字列を結合する式は，=CONCATENATE(C2,"/",D2,"/",E2)と等しい。

図－2　エクセルによる生年月日の数値データ化

	A	B	C	D	E	F
1	No	住所	全角文字数	半角文字数	分離した地区名	分離した番地
2	1	江見 12345	7	9	江見	12345
3	2	貝渚1－2－3	7	14	貝渚1－2－3	(※1)
4			=LEN(B3)	=LENB(B3)	=LEFT(B3,D3-C3)	=RIGHT(B3,C3-(D3-C3))

全角文字数(2バイト)と半角文字数(1バイト)の差を利用して，全角漢字の文字数を求める。

左から全角文字数分読み込めば地区名が分離される。
番地は，右から全角文字数を除いた文字数を読み込めば分離できる。
(※1)　貝渚1－2－3の様に全角ですべて入力されていると分離されない。
　　ASC関数を使用して全角（2バイト）の英数カナ文字を半角（1バイト）の文字に変換する。

	A	B	C	D
1		貝渚1－2－3		貝渚 1-2-3
2				=ASC(B2)

図－3　エクセルによる地区名と番地の分離

	A	B	C	D	E	F	G	H
1	No.	氏名	氏名の文字数	左から1文字を取り出し	左から2文字目	左から3文字目	左から4文字目	左から5文字目
2	1	千葉衛一郎	5	千	葉	衛	一	郎
3			=LEN(C2)	=LEFT(C2, 1)	=MID(C2, 2 ,1)	=MID(C2, 3, 1)	=MID(C2,4,1)	=MID(C2, 5, 1)

図－4　エクセルによる文字の抽出作業

3）個人情報の保護について

事前の協議の結果，個人識別情報のついた検診結果は市町の外に持ち出さず，市町から入手する情報は第三者が照合不可能なIDをつけた基本健康診査結果とした。また，作業は衛生研究所職員が各市町の庁舎内で実施し，突合結果は担当者のチェックを受けてから入手した。

（※）個人情報保護法では，「第二条　この法律において「個人情報」とは，生存する個人に関する情報であって，当該情報に含まれる氏名，生年月日その他の記述等により特定の個人を識別することができるもの（他の情報と容易に照合することができ，それにより特定の個人を識別することができることとなるものを含む。）をいう。」と規定されている。

4）突合作業手順の事前検討

突合作業は，旧市町の庁舎内の作業となるため作業方法の確認を行った。

検診受診者リストから承諾者リストに含まれる個人を抽出する作業をシミュレーションする目的で，アンケート調査のオプションとして栄養状況・運動状況解析結果を送付するために用意されたリストと承諾者リストを用いて抽出作業を行った。作業は，運動栄養解析結果送付リストと調査協力承諾者リストの二つのファイルをアクセスにインポートし，クエリでキーを指定し抽出作業を行った。その結果，複数項目をキーとすると，誤りが少なく，もれなく抽出できることが確認された（表－2）。旧天津小湊町の承諾者についても同じ結果が得られた。

表－2　運動栄養解析結果返送者リスト(10,740 人)と旧鴨川市承諾者(5,205 人)を用いた突合実験結果

クエリ抽出キー	データ有効者数(*1)	クエリ抽出数	間違い数	有効数	間違い/有効数%	対承諾者比
氏名	7,810	5,597	393	5,204	8%	1.000
氏名，地区名	7,796	5,231	29	5,202	1%	0.999
氏名，生年月日	7,605	5,187	2	5,185	2%	0.996
氏名，生年月日，地区名	7,595	5,185	2	5,183	0%	0.996
生年月日，地区名	7,596	5,271	87	5,184	2%	0.996
生年月日	10,014	8,999	3,814	5,185	74%	0.996
地区名	7,189	1,106,534				

(*1) 運動栄養解析結果返送者リストのそれぞれのデータのそろっている人数。

6．平成16年度突合作業

1）作業日数

事前の打ち合わせを含めて，平成16年８月11日から平成17年１月13日の間，旧鴨川市ふれあいセンター庁舎内10日，旧天津小湊町役場庁舎内５日の突合作業を行った。

2）当初作業方法

２市町３年度分の作業のうち平成16年度旧鴨川市について先行して実施し，作業方法の検証を行うこととした。また，旧市町における作業は表－3に従い実施した。

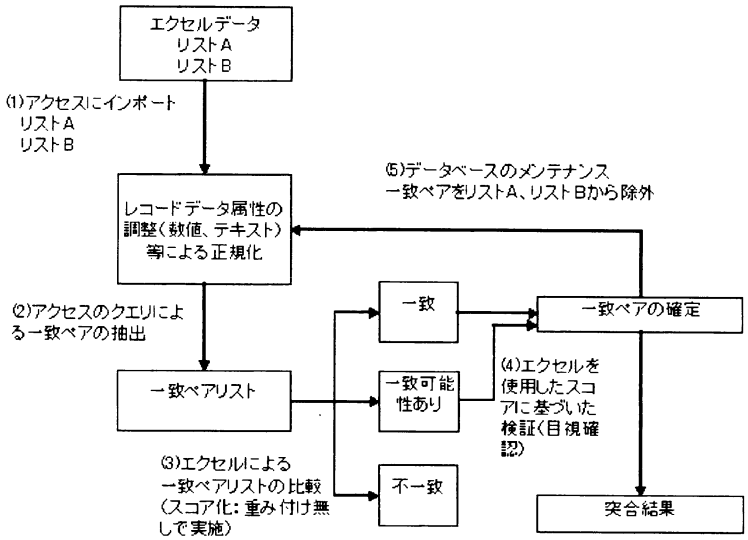
平成15，16年度総合検診結果２市町分について，それぞれ，総合検診受診者リストを作成し，氏名・住所・生年月日の情報を基に調査協力承諾者リストと突合を行った。昭和62年度データには，漢字情報がなく，氏名はカタカナのみ，承諾者の姓をカタカナ読みしたファイルを作成しカタカナ姓と生年月日で突合を行った。

アクセスによる抽出は，複数段階とした（図－5）。第１段階として氏名，地区名，生年月日，それぞれ全ての一致した者を採用することとした。第二段階として，氏名の１文字目と地区名，生年月日それぞれ全ての一致した者を採用し，第三段階目ではさらに基準を緩和することとした。各段階で抽出者の確認作業を行った（図－6）。しかし，昭和62年度検診結果は，前述の理由で別の取扱を決めた。まず，姓（カナ），生年月日による突合作業を行い，次に，姓（カナ１文字目），生年月日による突合作業を行うこととした。その結果については，全員について氏名（漢字）と氏名（カナ）の目視による確認作業を行うこととした。

表－3　旧鴨川市，旧天津小湊町における作業

1）旧市町担当者とのデータの受け渡し確認
2）作業用パソコンへのデータのコピー
3）エクセルによるデータの加工
　(1)個人識別情報部分の取り出し（受検番号,個人コード，氏名，住所，生年月日）
　(2)承諾者リストと整合性をとるためのデータの加工
4）アクセスによるレコードリンケージ
　(1)エクセルデータのインポート
　(2)抽出用クエリ作成　（氏名，地区名，生年月日を使用）
5）エクセルによるデータの照合と検証　氏名，地区名，生年月日の確認
6）照合結果　idリスト作成ファイルのデータディスクへの書き出し
7）作業用パソコン上のデータ消去
8）検診データと結合し個人情報除去後，市町の確認後データの受領

（※）作業は１日で終了しないので，退庁前に作業ファイルをディスクに保管し旧市町担当者に提出。パソコン上のデータを全て消去した後退庁した。次回の作業日に，保管を依頼したディスクを用いて作業を継続した。



図－5　突合作業フロー

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		照合項目								照合による得点				
2	NO	承諾氏名	承諾地区名	承諾番地	承諾生年月日	検診氏名	検診地区名	検診番地	検診生年月日	氏名	地区名	番地	生年月日	合計
3	1	##	##	##	##	##	##	##	##	1	1	1	1	4
4	2									=IF(EXACT(B3,F3)=TRUE,1,0)				=SUM(J3:M3)

マイクロソフトアクセスのクエリ抽出結果をエクセルにエクスポートし，照合項目の確認を行った。
図－6　エクセルを用いた一致ペアの比較

3）突合結果の検証

旧鴨川市平成16年度検診受信者に対する突合作業が終了し1,704名の突合作業が終了した時点で，作業方法の検証作業を実施した。11月17日旧鴨川市庁舎内でデータの再確認を実施した。

(1)　検証作業方法

承諾者リストと平成16年度検診受診者リストに対して表－4のクエリ抽出条件で抽出を行い，先に得られている平成１６年度分突合結果と比較を行った。

(2)　検証作業の結果

氏名全体と地区名，生年月日を組み合わせて抽出を行うと，誤って抽出されることはなかったが，抽出数が不足した。抽出基準をゆるめると，一人に対して複数が抽出され間違いが多数発生した。特に，住所（地区名）を用いた場合は，同一となる住民が多いため膨大な数が抽出されることになった。

以上のことから，複数の項目をもとに照合を行うことの正当性が確認された。(1)氏名，住所，生年月日の利用が確実性が高く有効であった。抽出数を増加させるために，(8)，(4)，(5)，(9)の条件で抽出を実施することとした。氏名，住所，生年月日のうち複数の項目が一致すれば候補にあげられるようにする。氏名の一部欠損による漏れをなるべくなくすように配慮を行うこととした。

表－4　検診受診者リストと承諾者リストでのクエリ抽出条件と抽出結果による作業手法の検証作業結果

No	クエリ抽出条件			抽出数			間違い
	氏名	住所(地区名)	生年月日	*1	*2	*3	
(1)	全体一致	一致	一致	1,419			なし
(2)	全体一致	一致		1,448			なし
(3)	全体一致		一致	1,433			なし
(4)		一致	一致	1,717			あり
(5)	全体一致			1,657			あり
(6)		一致		5,600,005			あり
(7)			一致	3,067			あり
(8)	一字一致	一致	一致	1,567	1,620	1,599	あり
(9)	一字一致	一致		18,616	29,446	6,102	あり

*1：氏名全体一致，または，１文字目一致，*2：氏名２文字目一致，*3：氏名３文字目一致
正解は，平成１６年度分として作業が終了した突合結果1,704人とした。

4）平成16年度突合作業結果

平成16年度旧鴨川市分の検証結果を受けて，平成15年度，16年度，昭和62年度分の検診データについて作業を行った。平成15年度は，旧鴨川市1,737名，旧天津小湊町556名。平成16年度分は，旧鴨川市1,701名，旧天津小湊町557名。昭和62年度分は，旧鴨川市検924名，旧天津小湊町416名の健診結果がアンケートIDにより分析可能となった。

5）データの受領方法

アンケートIDと検診データを突合し結果を記録したディスク(旧鴨川市MO，天津小湊町FDD)を提出し，個人識別情報が含まれないことの確認を受けた後に受領した。

7．突合作業で生じた問題点

平成16年度作業上で発生した問題点は，大きく2点となった。まず，第一は，承諾者データベースと健診結果の個人情報記載様式の違い。第二は，アンケートや承諾書と旧市町の公式記録との間での数や文字の記載違いであった。第一の記録様式の違いは，事前に想定していたものの他に，情報処理システムの文字コード等の違いが原因となった非表示や誤認識，さらに，昭和62年度の情報における独特なカタカナ表現（庄司：ショウジ→シヨウジ，小文字が使用されていない）などが存在した。第二は，記入時のミスと思われる名前や生年月日の違いが多数確認された。

8．まとめ

個人情報保護法が施行され，市町村から個人情報の提供を受けて，突合作業を実施することは難しくなると考えられる。今回は，市町の個人情報を持ち出さないために，作業は全て市町の庁舎内で実施し，持ち出すデータは個人情報の削除を担当者の確認を得た上で入手した。

また，疫学調査のためには，承諾者管理のために大がかりなデータベースを構築する必要が生じる。記載内容を確実にするために，承諾書には，氏名にはふりがなを書く欄をもうける，生年月日は昭和・平成□年□月□日のような形式で記入欄を用意して記入する形式が望まれる　さらに可能であれば，調査対象者抽出の段階で対象者に新たに番号を付け，その番号を基に郵送，承諾書の受付や検診等のデータを入手する手続きを，対象となる市町村との間で取り決めておくことが望まれる。

参考文献

1）相澤彰子他　異種データベース間でのレコード照合に関する研究動向　NII Journal No.8（2004.2）p43-51

2）柳川洋他，レコード・リンケージに関する基礎的研究　日本公衛誌　第18巻　第8号　p.487-493

3）Geoffrey R, Howe Use of Computerized Record Linkage in Cohort Studies. Epidemiologic Reviews 1998 Vol. 20 ; 1 : 112-121