資　料

国民健康・栄養調査の非協力者を同定するための国民生活基礎調査とのレコード・リンケージにおけるキー変数の組合せに関する検討

池田　奈由*　西　信雄*

目的　国民健康・栄養調査（以下，栄養調査）の非協力者を同定するためには，国民生活基礎調査（以下，基礎調査）から得られる親標本とのレコード・リンケージが必要である。レコード・リンケージは原則として，被調査者に割り振られた世帯番号等のキー変数により行われるが，誤った連結および非連結が発生する場合がある。本研究では，栄養調査の非協力者を同定するためのレコード・リンケージのキー変数の組合せについて検討した。

方法　1988～2015年（2012年除く）の基礎調査と栄養調査の個人データを用いて，栄養調査協力単位区における基礎調査協力者（栄養調査時点の0歳児を除く）と栄養調査協力者のレコード・リンケージを行った。キー変数の組合せには，「都道府県番号，地区番号，単位区番号，世帯番号，世帯員番号」（A），Aから世帯員番号を除き性別，出生年月または年齢を追加（B），Aに性別，出生年月または年齢を追加（C），BとCの二段階（D）の4通りを用いた。連結された協力者，連結されなかった基礎調査協力者（栄養調査非協力者とみなす），連結されなかった栄養調査協力者の3群に分類し，栄養調査協力者の連結率および基礎調査協力者の非連結率を，4通りの組合せの間で比較した。

結果　レコード・リンケージの対象となった基礎調査協力者は455,854人，栄養調査協力者は335,010人であった。調査年別の栄養調査協力者の連結率は，A（90％台後半），D（90％台前半），B（90％台前半），C（80％台）の順に高かった。Cに比べてAは8～14％ポイント，Bは5～10％ポイント高く，Bに比べてDは0.1～0.4％ポイント高かった。基礎調査協力者の非連結率は，C，B，D，Aの順に高く，Dでは1990年以前に20％台前半，1990年代に30％前後，2000年代に30％台～40％前後で推移した。

結論　キー変数の組合せにより，連結結果に差異があった。世帯員番号の変更や性別または出生年月，年齢の誤入力，同性の複産児の重複を考慮した二段階のレコード・リンケージにより，最も多くの栄養調査協力者を正確に連結できた。ただし，さらに世帯番号が変更された場合等への対応には限界がある。連結されなかった基礎調査協力者を栄養調査の非協力者とみなす際には，誤った連結結果が依然として存在する可能性があることに留意する必要がある。

Key words：国民健康・栄養調査，国民生活基礎調査，レコード・リンケージ，確定的リンケージ，非協力者

日本公衆衛生雑誌　2019; 66(4): 210–218.　doi:10.11236/jph.66.4_210

Ⅰ　緒　　言

国民健康・栄養調査（以下，栄養調査）は，健康増進法（平成14年法律第103号）に基づき，厚生労働省が毎年11月に実施する統計調査である。生活習慣調査票，栄養摂取状況調査票および身体状況調査票から構成され，国民の身体と生活習慣に関する状況を明らかにし，国民の健康増進を総合的に推進するために必要な基礎資料を得ることを目的としている[1]。健康日本21（第二次）が掲げる目標項目のデータソースとしても用いられており[2]，国民の健康に関する政策評価において重要性が極めて高い調査である[3]。

* 国立研究開発法人医薬基盤・健康・栄養研究所国際栄養情報センター
責任著者連絡先：〒162–8636 新宿区戸山1–23–1
国立研究開発法人医薬基盤・健康・栄養研究所国際栄養情報センター　池田奈由

栄養調査は，毎年6月および7月実施の国民生活基礎調査（以下，基礎調査）の後続調査として設計されている。基礎調査は，国民生活の基礎的事項を調査し，厚生労働行政の企画および運営に必要な基礎資料を得るとともに，栄養調査をはじめ各種調査の調査客体を抽出するための親標本を設定することを目的として，厚生労働省が実施する基幹統計調査である[4]。栄養調査の調査対象は，2012年と2016年に実施された拡大調査を例外として，基礎調査で設定された単位区（一単位区当たり20～30世帯）から層化無作為抽出した300単位区内の世帯および当該世帯の満1歳以上（11月1日現在）の世帯員である[1]。

近年，栄養調査における協力率の低さが問題となっている[5]。たとえば2015年の調査では，調査対象5,327世帯，調査実施3,507世帯と報告されており[1]，世帯単位の協力率は約66％であった。協力率の問題を検討する際には，栄養調査対象者のうち協力しなかった者を同定する必要がある。そのためには，栄養調査対象者のデータを基礎調査から抽出し，栄養調査とレコード・リンケージを行ったうえで，連結されなかった者を非協力者とすることが行われてきた[5]。原則的には，被調査者に割り振られる都道府県番号，地区番号，単位区番号，世帯番号，世帯員番号は基礎調査と栄養調査の間で共通であり，これらをキー変数としてレコード・リンケージを行うことが可能である。先行研究では，この方法でレコード・リンケージを行い，2003～2007年調査の協力者の特徴を検討した[5]。

一方で，他の先行研究で同様の方法を用いてレコード・リンケージを行ったところ，相当数のケースで性別と年齢が一致しないことが指摘されている[6]。その解釈として，いずれかの調査で性・年齢に誤入力があった同一人物，あるいは世帯番号または世帯員番号に変更があった別人が連結された可能性が考えられる。まず，栄養調査では調査票の回収時に調査員が自記式による記入内容を対面で確認する機会があるが[7]，その機会がなければ性別，出生年月または年齢に誤記入があっても訂正することができない。さらに，調査世帯名簿を作成する際，世帯番号は基礎調査と一致させることとなっている[8]。ただし，栄養調査は食事状況を調査するため，基礎調査では別世帯とされた複数の世帯について，調査日に食生活を共にしていれば同一世帯とみなして世帯番号を統一し，抹消した世帯番号の世帯員は，主となる世帯における世帯員の末尾番号に続くように変更することが認められていることなどから，世帯番号および世帯員番号が変更される場合がある（国民健康・栄養調査実施経験者からの情報による）。

1995年調査のデータを用いた先行研究では，都道府県番号，地区番号，単位区番号，世帯番号，性別，出生年月をキー変数とし，栄養調査協力者のほとんどが基礎調査のデータと連結された[8]。ただし，世帯員番号をキー変数に含めないことにより，世帯内に同性・同年齢の複産児などがいる場合，重複として処理され連結できないという問題点が残された[8]。また，2003～2007年調査のデータを用いた他の先行研究では，都道府県番号，地区番号，世帯番号，世帯員番号，性別，年齢をキー変数として，約50,000人の栄養調査協力者のうち約40,000人が連結された[9]。一方，基礎調査と栄養調査のレコード・リンケージにより得られたデータを活用した研究結果が発表されているが[10,11]，レコード・リンケージの詳細については記述されていない。

このように，基礎調査と栄養調査のレコード・リンケージにおいては，一部のキー変数の変更や誤入力等により誤った連結および非連結が発生し，非協力者が正しく同定されない場合がある。しかし，これまで基礎調査と栄養調査のレコード・リンケージの方法に焦点を当てた検討は行われておらず，キー変数について一致した見解は得られていない。そこで本稿では，栄養調査の非協力者を同定するためのレコード・リンケージのキー変数の組合せについて検討した。

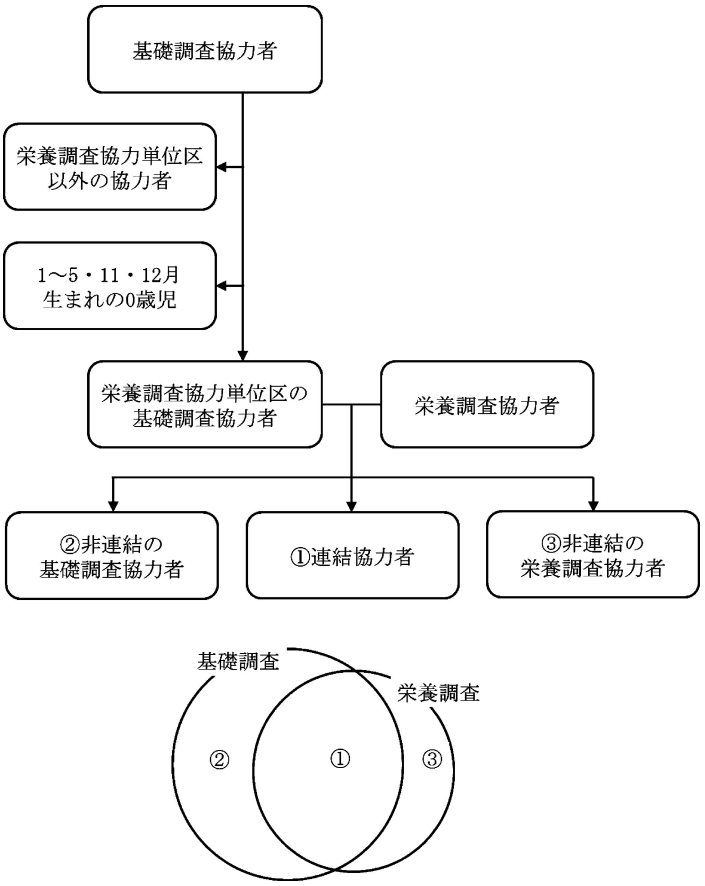
Ⅱ　研究方法

1.　データ

厚生労働省より，統計法（平成19年法律第53号）第33条の規定に基づき，栄養調査および基礎調査世帯票の調査票情報の提供を受けた。基礎調査の調査票情報は1986年調査から利用可能であるが，本分析では単位区による抽出に基づくレコード・リンケージが可能な1988～2015年調査のデータを用いた。なお，2012年栄養調査については，拡大調査として基礎調査を介さず国勢調査地区から標本が抽出されたため，本分析に含めなかった。

レコード・リンケージの対象は，栄養調査協力者全員のデータと，基礎調査協力者のうち栄養調査協力者が存在する単位区（栄養調査協力単位区）における者のデータとした。ただし，基礎調査協力者のうち1～5月または11～12月生まれの0歳児は，栄養調査時点で1歳未満であるため，レコード・リンケージの対象から除外した（図1）。栄養調査協力単位区を同定するため，毎年の栄養調査から都道府県番号，地区番号，単位区番号の組合せのリストを

図1　基礎調査と栄養調査のデータ準備とレコード・リンケージのフロー図と調査協力者の分類



作成した。このリストを同年の基礎調査の個人データに連結し，連結された者を栄養調査協力単位区における基礎調査協力者として抽出した。ただし，栄養調査のデータから調査対象単位区リストを作成したため，栄養調査の調査対象として抽出されたが栄養調査協力者が存在しなかった単位区があったとしても同定不可能であった。したがって，このような単位区における基礎調査協力者は栄養調査非協力者であるが，レコード・リンケージの対象外となった。

2.　レコード・リンケージ

一対一の確定的リンケージ（deterministic record linkage）により，栄養調査のデータを栄養調査協力単位区における基礎調査協力者の基礎調査のデータと連結した。キー変数の組合せとして，都道府県番号，地区番号，単位区番号，世帯番号，世帯員番号（A），都道府県番号，地区番号，単位区番号，世帯番号，性別，出生年月または年齢（B），都道府県番号，地区番号，単位区番号，世帯番号，世帯員番号，性別，出生年月または年齢（C）の3通りを用いた。まず，A，B，Cのそれぞれを用いた一段階のレコード・リンケージを行った。次に，二段階のレコード・リンケージとして，Bによるレコード・リンケージの後，連結されずに残った者に対してCによるレコード・リンケージを行った（D）。なお，Bを含むレコード・リンケージ（BとD）では，世帯員番号の除外により発生した重複を予め取り置き，それらを含まないデータに対してレコード・リンケージを行った。

出生年月または年齢については，B，C，Dのそれぞれにおいて，まず出生元号，出生年，出生月の組合せである出生年月を用いたレコード・リンケージを行い，連結されずに残った者を対象に年齢を用いたレコード・リンケージを行った。さらに，出生年月または年齢で連結されなかった場合，特に基礎調査から得られた出生月が6〜10月の者については，基礎調査から栄養調査までの間で誕生日を迎えるため，基礎調査での年齢に1歳を加えた値を栄養調査での年齢に照合したレコード・リンケージを追加した。なお，2000〜2010年については，栄養調査のデータに出生年月に関する変数がなかったため，出生年月を用いたレコード・リンケージを省略した。

3.　分析方法

レコード・リンケージの連結結果により，分析対象を連結された協力者（①），連結されなかった基礎調査協力者（②）および連結されなかった栄養調査協力者（③）の3群に分類し（図1），②を栄養

調査非協力者とみなした。栄養調査協力者の連結率（①/（①＋③）×100）および基礎調査協力者の非連結率（②/（①＋②）×100）を計算し，A～Dの間で比較した。

Ⅲ　研究結果

1.　栄養調査協力者の連結率

レコード・リンケージの対象となった栄養調査協力者は合計335,010人で，調査年別では8,583人（2015年）～18,161人（1988年）であった（表1）。一段階のレコード・リンケージの結果，栄養調査協力者の連結率は，Aで94.5％（2004年）～98.3％（1995年），Bで88.8％（1999年）～94.6％（1995年），Cで82.6％（1996年）～89.0％（2015年）であった（表1）。相互に比較すると，AよりもBの方が2.2％ポイント（2013年）～6.3％ポイント（1999年）低く，BよりもCの方が5.1％ポイント（2015年）～10.0％（1995年）低く，AよりもCの方が8.1％ポイント（2015年）～14.0％ポイント（1996年）低かった。二段階のレコード・リンケージ（D）では，一段階目のBの後，二段階目のCで0.2％（2004年）～0.5％（2010年）が追加的に連結され，合計で栄養調査協力者の連結率は89.0％（1999年）～94.8％（1995年）であった（表1）。

Dについて出生年月と年齢を用いた連結結果の内訳を見ると，栄養調査データに出生年月がある

表1　栄養調査協力者のうち基礎調査データと連結された者の人数（％）

調査年	協力者	キー変数の組合せ			
		A	B	C	D
1988	18,161	17,506(96.4)	16,778(92.4)	15,512(85.4)	16,826(92.6)
1989	16,657	16,001(96.1)	15,234(91.5)	13,943(83.7)	15,289(91.8)
1990	17,986	17,413(96.8)	16,677(92.7)	15,029(83.6)	16,722(93.0)
1991	16,711	16,250(97.2)	15,652(93.7)	14,356(85.9)	15,682(93.8)
1992	15,694	15,320(97.6)	14,762(94.1)	13,470(85.8)	14,806(94.3)
1993	15,782	15,353(97.3)	14,636(92.7)	13,411(85.0)	14,685(93.0)
1994	14,546	13,925(95.7)	13,282(91.3)	12,125(83.4)	13,317(91.6)
1995	14,240	14,002(98.3)	13,465(94.6)	12,038(84.5)	13,498(94.8)
1996	14,019	13,546(96.6)	12,711(90.7)	11,580(82.6)	12,744(90.9)
1997	13,289	12,841(96.6)	12,171(91.6)	11,236(84.6)	12,205(91.8)
1998	14,159	13,696(96.7)	13,119(92.7)	11,868(83.8)	13,151(92.9)
1999	12,763	12,130(95.0)	11,328(88.8)	10,553(82.7)	11,355(89.0)
2000	12,271	11,808(96.2)	11,224(91.5)	10,299(83.9)	11,257(91.7)
2001	12,481	11,938(95.6)	11,504(92.2)	10,573(84.7)	11,536(92.4)
2002	11,491	10,959(95.4)	10,444(90.9)	9,686(84.3)	10,468(91.1)
2003	11,628	11,171(96.1)	10,706(92.1)	9,950(85.6)	10,753(92.5)
2004	9,484	8,959(94.5)	8,706(91.8)	8,087(85.3)	8,717(91.9)
2005	9,561	9,224(96.5)	8,948(93.6)	8,333(87.2)	8,968(93.8)
2006	9,923	9,471(95.4)	9,057(91.3)	8,405(84.7)	9,086(91.6)
2007	9,611	9,199(95.7)	8,847(92.1)	8,209(85.4)	8,883(92.4)
2008	9,886	9,449(95.6)	9,053(91.6)	8,386(84.8)	9,072(91.8)
2009	9,942	9,642(97.0)	9,336(93.9)	8,677(87.3)	9,359(94.1)
2010	9,635	9,222(95.7)	8,978(93.2)	8,309(86.2)	9,018(93.6)
2011	8,761	8,475(96.7)	8,126(92.8)	7,583(86.6)	8,140(92.9)
2012			分析対象外		
2013	8,619	8,233(95.5)	8,044(93.3)	7,491(86.9)	8,065(93.6)
2014	9,127	8,853(97.0)	8,521(93.4)	7,986(87.5)	8,553(93.7)
2015	8,583	8,327(97.0)	8,071(94.0)	7,635(89.0)	8,085(94.2)

A：都道府県番号，地区番号，単位区番号，世帯番号，世帯員番号
B：都道府県番号，地区番号，単位区番号，世帯番号，性別，出生年月または年齢または基礎調査の年齢に1歳を加えた年齢（6～10月生まれのみ）
C：都道府県番号，地区番号，単位区番号，世帯番号，世帯員番号，性別，出生年月または年齢または基礎調査の年齢に1歳を加えた年齢（6～10月生まれのみ）
D：Bによるレコード・リンケージの後，連結されなかった者に対してCによるレコード・リンケージを追加

1988～1999年と2011～2015年では，出生年月で85.9％（1999年）～93.4％（1995年）が連結され，年齢または基礎調査での年齢に1歳を加えた年齢で0.9％（2011年）～6.6％（2015年）が追加的に連結された（表2）。一方，栄養調査データに出生年月がない2000～2010年では，年齢で54.9％（2000年）～56.9％（2009年）が連結され，基礎調査での年齢に1歳を加えた年齢で35.9％（2006年）～37.3％（2010年）が追加的に連結された（表2）。栄養調査協力者の非連結率は，5.2％（1995年）～11.0％（1999年）であった。

2. 基礎調査協力者の非連結率

レコード・リンケージの対象となった基礎調査協力者は合計455,854人で，調査年別では12,881人（2013年）～22,053人（1988年）であった（表3）。基礎調査協力者の非連結率は，いずれの調査年においてもAで最も低くCで最も高く，AとCの差は5.1％ポイント（2015年）～10.9％ポイント（1995年）であった（表3）。BはCよりも3.2％ポイント（2015年）～7.9％ポイント（1995年）低く，Dよりも0.1％ポイント（2011年）～0.3％ポイント（2003年）低かった。A～Dのいずれにおいても，非連結率は1988年から2015年にかけて緩やかに上昇した。例えばDでは，1990年以前は20％台前半で，1990年代に30％前後を推移し，2000年代に入ると30％台から40％前後を推移した。

Ⅳ 考　　察

本稿では，栄養調査における協力率に関する検討の一環として，基礎調査と栄養調査とのレコード・リンケージの方法と非協力者の同定について検討した。その結果，キー変数の組合せによって栄養調査

表2　栄養調査協力者のうち二段階のレコード・リンケージ（D）で出生年月または年齢により基礎調査データと連結された者と連結されなかった者の人数（％）

調査年	協力者	出生年月または年齢による連結			非連結
		出生年月	年　齢	年齢＋1歳[a)]	
1988	18,161	16,518(91.0)	199(1.1)	109(0.6)	1,335(7.4)
1989	16,657	15,045(90.3)	158(0.9)	86(0.5)	1,368(8.2)
1990	17,986	16,454(91.5)	176(1.0)	92(0.5)	1,264(7.0)
1991	16,711	15,439(92.4)	137(0.8)	106(0.6)	1,029(6.2)
1992	15,694	14,574(92.9)	147(0.9)	85(0.5)	888(5.7)
1993	15,782	14,460(91.6)	138(0.9)	87(0.6)	1,097(7.0)
1994	14,546	13,096(90.0)	137(0.9)	84(0.6)	1,229(8.4)
1995	14,240	13,303(93.4)	133(0.9)	62(0.4)	742(5.2)
1996	14,019	12,425(88.6)	181(1.3)	138(1.0)	1,275(9.1)
1997	13,289	11,935(89.8)	177(1.3)	93(0.7)	1,084(8.2)
1998	14,159	12,720(89.8)	253(1.8)	178(1.3)	1,008(7.1)
1999	12,763	10,966(85.9)	231(1.8)	158(1.2)	1,408(11.0)
2000	12,271	0(0.0)	6,731(54.9)	4,526(36.9)	1,014(8.3)
2001	12,481	0(0.0)	6,981(55.9)	4,555(36.5)	945(7.6)
2002	11,491	0(0.0)	6,329(55.1)	4,139(36.0)	1,023(8.9)
2003	11,628	0(0.0)	6,497(55.9)	4,256(36.6)	875(7.5)
2004	9,484	0(0.0)	5,255(55.4)	3,462(36.5)	767(8.1)
2005	9,561	0(0.0)	5,417(56.7)	3,551(37.1)	593(6.2)
2006	9,923	0(0.0)	5,527(55.7)	3,559(35.9)	837(8.4)
2007	9,611	0(0.0)	5,314(55.3)	3,569(37.1)	728(7.6)
2008	9,886	0(0.0)	5,481(55.4)	3,591(36.3)	814(8.2)
2009	9,942	0(0.0)	5,661(56.9)	3,698(37.2)	583(5.9)
2010	9,635	0(0.0)	5,427(56.3)	3,591(37.3)	617(6.4)
2011	8,761	8,059(92.0)	56(0.6)	25(0.3)	621(7.1)
2012			分析対象外		
2013	8,619	7,729(89.7)	217(2.5)	119(1.4)	554(6.4)
2014	9,127	8,071(88.4)	305(3.3)	177(1.9)	574(6.3)
2015	8,583	7,518(87.6)	334(3.9)	233(2.7)	498(5.8)

[a)] 6～10月生まれについては，基礎調査の年齢に1歳を加えた年齢と栄養調査の年齢を照合

協力者の連結率と基礎調査協力者の非連結率に差異が生じることが初めて示された。

被調査者に一意的に割り振られた都道府県番号，地区番号，単位区番号，世帯番号，世帯員番号（A）による調査協力者の連結率は90％台後半であったが，キー変数に性別，出生年月または年齢を加えると80％台に低下した（C）。これは，都道府県番号から世帯員番号までは一致したものの，性別または出生年月，年齢が異なるペアが存在することを示している。その背景として，いずれかの調査で性別または出生年月，年齢に誤入力があった場合と，栄養調査で世帯番号または世帯員番号が変更された場合が考えられる。前者の場合はAが正しくCは誤りで，後者の場合はAは誤りでCが正しい。

次に，Cから世帯員番号を除外し，栄養調査での世帯員番号の変更を許容したBでは，5～10％ポイントの連結率の向上が見られた。この変化分は，世帯員番号が変更されたため世帯員番号が一致せずCで誤って連結されなかった同一人物の集団と，世帯内で同性・同年齢の他の世帯員と重複しておりBで誤って連結されなかった複産児等の集団との間の差である。なお，先行研究でも，キー変数に世帯員番号を含まないレコード・リンケージを行ったところ，複産児を連結できなかったことに言及している[8]。ただし，この集団の規模は小さく，連結率の上昇分の大半は世帯員番号に変更のあった者がB

表3　栄養調査協力単位区における基礎調査協力者のうち連結されなかった者の人数（％）

調査年	協力者	キー変数の組合せ			
		A	B	C	D
1988	22,053	4,547(20.6)	5,275(23.9)	6,541(29.7)	5,227(23.7)
1989	20,331	4,330(21.3)	5,097(25.1)	6,388(31.4)	5,042(24.8)
1990	21,855	4,442(20.3)	5,178(23.7)	6,826(31.2)	5,133(23.5)
1991	21,111	4,861(23.0)	5,459(25.9)	6,755(32.0)	5,429(25.7)
1992	19,806	4,486(22.6)	5,044(25.5)	6,336(32.0)	5,000(25.2)
1993	20,403	5,050(24.8)	5,767(28.3)	6,992(34.3)	5,718(28.0)
1994	19,238	5,313(27.6)	5,956(31.0)	7,113(37.0)	5,921(30.8)
1995	17,958	3,956(22.0)	4,493(25.0)	5,920(33.0)	4,460(24.8)
1996	18,437	4,891(26.5)	5,726(31.1)	6,857(37.2)	5,693(30.9)
1997	17,506	4,665(26.6)	5,335(30.5)	6,270(35.8)	5,301(30.3)
1998	18,168	4,472(24.6)	5,049(27.8)	6,300(34.7)	5,017(27.6)
1999	16,979	4,849(28.6)	5,651(33.3)	6,426(37.8)	5,624(33.1)
2000	16,730	4,922(29.4)	5,506(32.9)	6,431(38.4)	5,473(32.7)
2001	16,950	5,012(29.6)	5,446(32.1)	6,377(37.6)	5,414(31.9)
2002	16,628	5,669(34.1)	6,184(37.2)	6,942(41.7)	6,160(37.0)
2003	15,970	4,799(30.1)	5,264(33.0)	6,020(37.7)	5,217(32.7)
2004	13,794	4,835(35.1)	5,088(36.9)	5,707(41.4)	5,077(36.8)
2005	14,587	5,363(36.8)	5,639(38.7)	6,254(42.9)	5,619(38.5)
2006	15,024	5,553(37.0)	5,967(39.7)	6,619(44.1)	5,938(39.5)
2007	14,334	5,135(35.8)	5,487(38.3)	6,125(42.7)	5,451(38.0)
2008	14,684	5,235(35.7)	5,631(38.3)	6,298(42.9)	5,612(38.2)
2009	14,892	5,250(35.3)	5,556(37.3)	6,215(41.7)	5,533(37.2)
2010	13,990	4,768(34.1)	5,012(35.8)	5,681(40.6)	4,972(35.5)
2011	14,179	5,704(40.2)	6,053(42.7)	6,596(46.5)	6,039(42.6)
2012			分析対象外		
2013	12,881	4,648(36.1)	4,837(37.6)	5,390(41.8)	4,816(37.4)
2014	13,822	4,969(35.9)	5,301(38.4)	5,836(42.2)	5,269(38.1)
2015	13,544	5,217(38.5)	5,473(40.4)	5,909(43.6)	5,459(40.3)

A：都道府県番号，地区番号，単位区番号，世帯番号，世帯員番号
B：都道府県番号，地区番号，単位区番号，世帯番号，性別，出生年月または年齢または基礎調査の年齢に1歳を加えた年齢（6～10月生まれのみ）
C：都道府県番号，地区番号，単位区番号，世帯番号，世帯員番号，性別，出生年月または年齢または基礎調査の年齢に1歳を加えた年齢（6～10月生まれのみ）
D：Bによるレコード・リンケージの後，連結されなかった者に対してCによるレコード・リンケージを追加

で正しく連結されたことによるものであると考えられる。このように，キー変数から世帯員番号を除外して世帯員番号の変更を考慮すると，レコード・リンケージの連結結果が向上することが分かる。

さらに，Dの二段階のレコード・リンケージでは，世帯内に同性・同年齢の他の世帯員が存在し，Bでは重複として連結されなかった者を，Cで補足的に連結した。BとDの間の連結率の差はわずかであったが，自動化できれば二段階のレコード・リンケージを行う手間もさほど大きくなく，複産児等も取りこぼさずに分析を行うことができる。このように，今回検討した4通りのキー変数の組合せのうち，Dが最も多くの栄養調査協力者と基礎調査協力者を正しく連結した。非協力者の同定のみならず種々の分析においては，BとCの二段階のレコード・リンケージにより連結されたデータを用いることが推奨される。なお，複産児を除外した分析では，Bの一段階のレコード・リンケージだけでよい。

栄養調査データに出生年月のある調査年において，出生年月を含むキー変数では連結されなかったが，出生年月の代わりに年齢を用いると連結される場合があった。このことから，出生年月と年齢のいずれかに誤入力がある場合に備えて一つずつキー変数に用いることにより，連結率が上昇することが分かる。また，同年の基礎調査と栄養調査との間で誕生日を迎える者については，基礎調査での年齢に1歳を加えた値と栄養調査での年齢を照合することによって連結可能であり，特に出生年月のない調査年において相当数に上ることが示された。

連結されなかった基礎調査協力者（図1の②）は，栄養調査に協力しなかった者と，栄養調査に協力したがキー変数が一致せず誤って連結されなかった者のいずれかである。後者の集団の大きさは，たかだか連結されなかった栄養調査協力者の数である。連結されなかった栄養調査協力者の中には基礎調査に協力せず栄養調査のみ協力した者が含まれる可能性はあるが，全員が基礎調査データと連結された場合，基礎調査協力者の非連結率は，Dによる値よりも3.7％ポイント（2015年）～8.3％ポイント（1999年）低かった。このように，連結されなかった基礎調査協力者を栄養調査の非協力者とみなす際には，誤って連結されなかった者が含まれている可能性に留意して分析結果を解釈する必要がある。

本研究の制約として，3点が挙げられる。1点目として，栄養調査の調査対象として抽出されたが協力しなかった単位区における基礎調査協力者は，レコード・リンケージで同定できないため分析対象外とした。ほぼ毎年，こうした単位区が少数ながら存在する可能性に留意する必要がある。2点目として，連結されなかった栄養調査協力者（図1の③）は，基礎調査とキー変数が一致しなかった場合と，基礎調査には協力しなかったものの何らかの理由で栄養調査のみ協力した場合のいずれかに当てはまると考えられる。前者については，世帯員番号以外の世帯番号，性別または出生年月（年齢）をキー変数から除外することによって連結可能な場合もあるが，同時に相当数の誤った連結も発生して確認作業が煩雑になるため，そのまま連結されない栄養調査協力者とみなした。後者については，単位区によっては事情があってやむを得ず調査対象者以外に調査を依頼した可能性があるが，データ上でそれを確認することは不可能である。3点目として，キー変数に世帯員番号を含むレコード・リンケージ（A，B，D）では，世帯内に同性・同年齢の複数の世帯員が存在する場合，世帯員番号に変更がないという仮定を置いているが，それがすべてのペアについて当てはまるかどうかをデータ上で確認することはほぼ不可能である。

なお，本研究では二段階のレコード・リンケージ（D）において，一段階目でB，二段階目でCを用いたが，逆にまずCですべてのキー変数を用いてレコード・リンケージを行い，連結されずに残った者についてBで世帯員番号を外し，条件を緩めたレコード・リンケージを行う方法もある[12]。しかし，栄養調査協力者の連結率のDとの差は約0.1％ポイント以下であったため，本分析ではDによる結果のみを示した。

Ⅴ　結　　語

基礎調査と栄養調査のレコード・リンケージにおいては，キー変数の組合せに注意する必要がある。まず，栄養調査における世帯員番号の変更を考慮して，都道府県番号，地区番号，単位区番号，世帯番号，性別，出生年月または年齢をキー変数とするレコード・リンケージを行った後，連結されなかった同性の複産児らについて世帯員番号をキー変数に加えて補足的に連結する二段階のレコード・リンケージにより，最も多くの栄養調査協力者と基礎調査協力者を正しく連結することが可能である。ただし，以上の手順で連結されなかった基礎調査協力者に基づき栄養調査の非協力者の特徴について検討する際には，誤って連結されなかった基礎調査協力者が依然として含まれる可能性に留意する必要がある。

本研究は，平成27～29年度科学研究費助成事業（基盤研究（C）（一般））「国民健康・栄養調査における非協力

者バイアス修正のための統計手法の開発と応用」15K08762（研究代表者：池田奈由）と平成30～32年度科学研究費助成事業（基盤研究（B）（一般））「非感染性疾患関連要因の推移と格差に関する大規模保健統計データの時空間的統合解析」18H03063（研究代表者：池田奈由）の一環として実施した。

本研究に関連し開示すべきCOI状態はない。

（受付　2018.11. 2）
（採用　2019. 1.23）

文　献

1）国立研究開発法人医薬基盤・健康・栄養研究所．国民健康・栄養の現状―平成27年厚生労働省国民健康・栄養調査報告より―．東京：第一出版．2018.
2）国立研究開発法人医薬基盤・健康・栄養研究所．健康日本21（第二次）分析評価事業．http://www.nibiohn.go.jp/eiken/kenkounippon21/index.html（2018年6月14日アクセス可能）.
3）Ikeda N, Takimoto H, Imai S, et al. Data resource profile: The Japan National Health and Nutrition Survey (NHNS). Int J Epidemiol 2015; 44: 1842–9.
4）厚生労働省政策統括官（統計・情報政策担当）．平成27年　国民生活基礎調査．東京：厚生労働統計協会．2017.
5）西　信雄，中出麻紀子，猿倉薫子，他．国民健康・栄養調査の協力率とその関連要因．厚生の指標 2012; 59(4): 10–15.
6）安藤雄一，青山　旬，尾崎哲則，他．歯科疾患実態調査の協力率に関する検討：平成23年歯科疾患実態調査の協力者は大半が国民健康・栄養調査における血液検査の協力者であった．日本公衆衛生雑誌 2016; 63: 319–324.
7）Ikeda N, Okuda N, Tsubota-Utsugi M, et al. Association of energy intake with the lack of in-person review of household dietary records: Analysis of Japan National Health and Nutrition Surveys from 1997 to 2011. J Epidemiol 2016; 26: 84–91.
8）橋本修二，川戸美由紀，松村康弘，他．保健統計におけるレコードリンケージの実施可能性．厚生の指標 2001; 48(11): 1–5.
9）Fukuda Y, Hiyoshi A. Associations of household expenditure and marital status with cardiovascular risk factors in Japanese adults: analysis of nationally representative surveys. J Epidemiol 2013; 23: 21–7.
10）Kadota A, Okuda N, Ohkubo T, et al. The National Integrated Project for Prospective Observation of Noncommunicable Disease and its trends in the aged 2010 (NIPPON DATA2010): Objectives, design, and population characteristics. J Epidemiol 2018; 28: S2–S9.
11）Tabuchi T, Nakamura M, Nakayama T, et al. Tobacco price increase and smoking cessation in Japan, a developed country with affordable tobacco: A national population-based observational study. J Epidemiol 2016; 26: 14–21.
12）Dusetzina S, Tyree S, Meyer A, et al. Linking data for health services research: A framework and instructional guide. (Prepared by the University of North Caroline at Chapel Hill under Contract No. 290–2010–000141.) AHRQ Publication No. 14–EHC033–EF. Rockville, MD: Agency for Healthcare Research and Quality. 2014.

Key variable combinations for identifying non-participants in the Japan National Health and Nutrition Survey through record linkage with the Comprehensive Survey of Living Conditions

Nayu IKEDA* and Nobuo NISHI*

Key words: National Health and Nutrition Survey, Comprehensive Survey of Living Conditions, record linkage, deterministic record linkage, nonparticipants

Objectives The identification of non-participants in the Japan National Health and Nutrition Survey (NHNS) requires record linkage with its master sample from the Comprehensive Survey of Living Conditions (CSLC). In principle, we can merge individual records between the two surveys by using key identifiers including household ID, but false matches and nonmatches can occur. We examined combinations of key variables for improving record linkage to identify nonparticipants in the NHNS.

Methods We used individual-level data from the NHNS and the CSLC from 1988 to 2015 (except 2012). We extracted from CSLC data individuals in participating unit blocks in the NHNS to merge records between the two surveys. We used four combinations of key variables: prefecture ID, census enumeration district ID, unit block ID, household ID, and household member ID (A); household member ID in A was replaced with sex and birth year and month or age (B); sex and birth year and month or age were added to A (C); two-stage linkage of B and C (D). We classified a sample of individuals into matched participants, unmatched NHNS participants, and unmatched CSLC participants (a proxy for nonparticipants). We compared the percentages of matched NHNS participants and unmatched CSLC participants across the four combinations of key variables.

Results We obtained a sample of 455,854 participants from the CSLC and 335,010 from the NHNS. The percentage of matched NHNS participants was highest in A (the upper 90%), followed by D (the lower 90%), B (the lower 90%), and C (the 80%). Compared to C, the percentage of matched NHNS participants was higher by 8–14 percentage points in A and 5–10 percentage points in B. Compared to B, it was higher by 0.1–0.4 percentage points in D. The percentage of unmatched CSLC participants was highest in C, followed by B, D, and A. The percentage of unmatched CSLC participants increased in D from the 20% level in the late 1980s to around 30% in the 1990s and stayed between the 30% level and the lower 40% level in the 2000s.

Conclusion The highest percentage of accurate matches of NHNS participants was obtained by considering changes in household member ID and incorrect entries on sex and birth year/month and age, and same-sex multiple births. However, there are limitations in handling unmatched participants due to changes in household ID or other reasons. It is therefore necessary to consider the possibility of false nonmatches included in unmatched CSLC participants in regarding them as non-participants in the NHNS.

* International Center for Nutrition and Information, National Institute of Health and Nutrition, National Institutes of Biomedical Innovation, Health and Nutrition