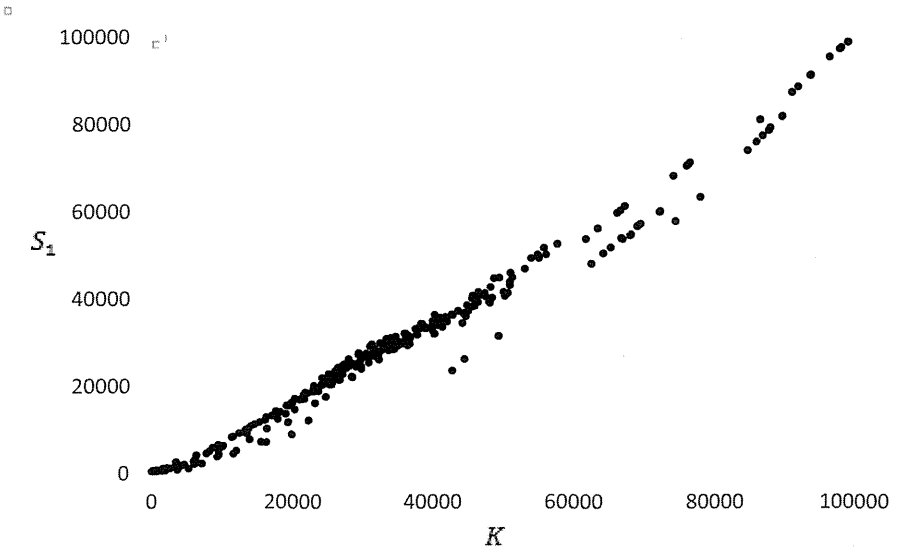


図1　分類数(K)とユニークセル数(S_1)の関係

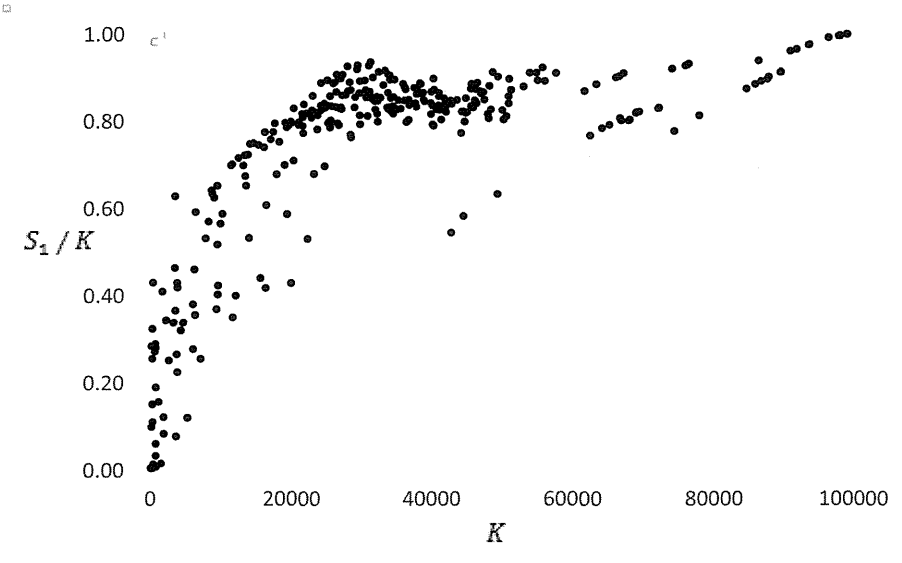


図2　分類数(K)と分類数に占めるユニークセル数の割合(S_1/K)の関係

平成25年度厚生労働科学研究費補助金(循環器疾患・糖尿病等生活習慣病対策総合研究事業)
「追跡終了後コホート研究を用いた共通化データベース基盤整備とその活用に関する研究」
分担研究報告書

Japan Arteriosclerosis Longitudinal Study (JALS)データにおける一意性の検討

研究分担者　大橋 靖雄　東京大学大学院医学系研究科
研究協力者　原田亜紀子　東京大学大学院医学系研究科
研究協力者　田島 里華　東京大学大学院医学系研究科

研究要旨

JALSのデータを用いて、死亡者の個人特定可能性について検討したところ、地域の情報が特に個人の特定に繋がりやすいことが明らかとなった。さらに参加コホートが属する市町村の死亡集計表をもとにした約30万の死亡者を対象に検討結果では、少数セルの発生は40～50歳台で多くみられていたが、人口30,000人以上の市町村では総死亡においては、年齢階級（10歳階級）別で少数セルの発生はみられなかった。集計対象を60歳以上に限定した場合では、5,000人以上の市町村であれば総死亡、循環器疾患死亡ともに少数セルの発生を抑えられる可能性が考えられた。

A. 目的

Japan Arteriosclerosis Longitudinal Study (JALS) 対象者で死亡したものにつき、特定の変数による個人特定可能性を検討する。さらに、JALS参加コホートが属する市町村の総死亡、循環器疾患死亡について、男女別に10歳階級別で死亡者数を集計した場合に、どのような条件において少数例（セル）が発生するかを検討する。

B. 方法

Ⅰ. JALS対象者における死亡者に関する検討

JALS対象者（年齢は40～89歳に限定）のうち、各コホートが住基情報等をもとに2010年度までに特定した死亡者4,858人について、特定の変数による個人特定可能性を検証した。

Ⅱ. JALS対象地域の死亡数と市町村規模の検討

JALS参加コホートが属する市町村について、年齢階級別の死亡集計表をもとに、1999年から2010年までの12年間で市町村合併のなかった88市町村　40歳以上の死亡者295,846例を分析対象に、市町村規模と特定可能性の検討を行った。40歳以上の集団で検討した場合と死亡数が多い60歳以上に限定した場合を検討した。

C. 結果

Ⅰ. JALS対象者における死亡者に関する検討

1. 単変数による個人特定可能性の検討

1-1. 性別

男性2,967人（61.1%）、女性1,891人（38.9%）であり、性別のみからは個人は特定されない。

1-2. ベースライン時年齢

図1より、ベースライン時年齢ごとの人数は最少でも6人であり、この変数のみにより個人を特定されない。

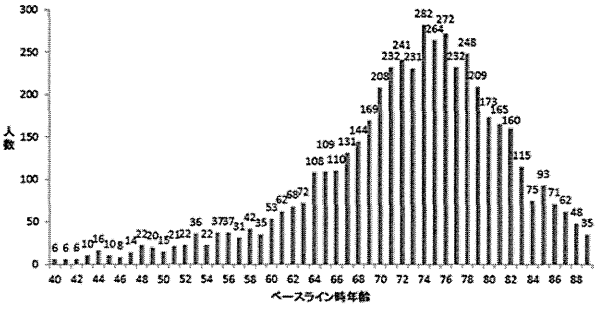


図1. ベースライン時年齢ごとの人数分布

1-3. 死亡時年齢

図2と表1より、死亡時年齢が40代前半もしくは90代後半の場合、死亡時年齢のみから個人を特定できる可能性は高くなる。

表 1. 死亡者数が 5 人以下であった死亡時年齢

死亡時年齢	人数
44	3
45	2
96	5
98	2
99	1

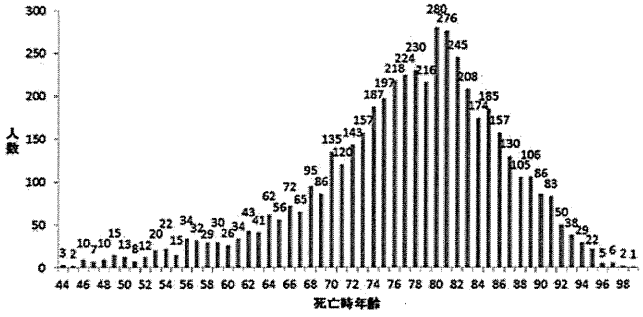


図 2. 死亡時年齢の分布

1-4. 地域（コホート・市町村）

コホート単位で死亡者を集計したところ、死亡者数が 20 名以下のコホートが 2 つ（12 人と 17 人）存在した。またコホート単位に比べより詳細な地域コード（市町村）ごとにも検討したところ、死亡者数が 1 名のみ、10 名以下、20 名以下の地域がそれぞれ 3 か所、9 か所、18 か所存在していた（表 2）。地域を公表する際にはコホート単位にとどめることにより、地域情報のみからの個人特定可能性を下げることができると考えられる。

表 2. 死亡者数が 20 人以下となる地域

地域コード	人数
33-3	1
13-11	1
33-14	1
33-6	2
33-2	3
33-1	4
33-13	7
33-4	8
2-5	9
3-17	11
21-1	12
3-13	12
3-16	13
29-1	17
2-1	17
3-11	18
1-2	19
3-6	19

*地域コード：コホート番号－コホート内の市町村通番

1-5. 死亡年月日

死亡年月日ごとの人数を集計した結果を表 3 に示した。全死亡者 4,858 人のうち、21.2%にあたる 1,028 人はユニークな死亡年月日を有していた（表 3）。また、同一死亡年月日を持つ人が 5 人以下である場合がほとんどであるため、この変数は特定性が高いと考えられた。死亡情報を死亡年月までに丸めて集計した結果を表 4 に示した。死亡年月までの丸めを行っても、個人が一意に特定されるパターンが 10 個存在し、これ以外にも同一死亡年月を持つ人数が 10 人以下のパターンも複数存在していた。よって死亡時点を公表する際には、死亡年まで丸めて公表するなど、データを粗くする必要性が考えられた。

表 3. 死亡年月日の特異性の検討

同一死亡日を持つ人数（人）	死亡日のパターン数	合計人数（人）	割合（%）
1	1028	1028	21.2
2	648	1296	26.7
3	361	1083	22.3
4	160	640	13.2
5	69	345	7.1
6	27	162	3.3
7	15	105	2.2
8	8	64	1.3
9	5	45	0.9
10	2	20	0.4
11	4	44	0.9
12	1	12	0.2
14	1	14	0.3

表 4. 死亡年月の特異性の検討

同一死亡年月を持つ人数（人）	死亡年月のパターン数
1	10
2	7
3	3
4	5
5	5
6	1
7	1
8	2
9	1

2.2 変数による個人特定可能性の検討

2-1. 性とベースライン時年齢

表 5 より、40 代の対象者においては、性別との組み合わせにより個人の特定可能性が高いが、ベースライン時年齢を 5 歳刻みのカテゴリ変数にすることなどで対処可能であると考えられる。

表 5. 人数が 5 人以下となる性・ベースライン年齢の組み合わせ

性	ベースライン時年齢	人数
男性	40	2
	41	2
	42	4
	47	4
女性	40	4
	41	4
	42	2
	43	2
	45	4
	46	1
	51	4

2-2. 性と死亡時年齢

表 6 より、男性では死亡時年齢が 40 代と 90 代、女性では 40, 50 代と 90 代では、性別と死亡時年齢を組み合わせることで個人が特定される可能性が高くなり、一意に特定できるパターンも 2 つ存在する。

表 6. 人数が 5 人以下となる性・死亡時年齢の組み合わせ

性別	死亡時年齢	人数
男性	44	2
	45	2
	47	4
	48	5
	96	3
	97	4
女性	44	1
	46	4
	47	3
	48	5
	50	4
	51	2
	52	5
	55	5
	96	2
	97	2
	98	2
	99	1

2-3. 性と地域

地域情報の公開をコホートレベルにとどめ、性別と組み合わせた場合、合計人数が 5 以下となるパターンは、女性に関して、2 人のみという地域が 1 か所存在した。

2-4. ベースライン時年齢と死亡時年齢

ベースライン時と死亡時の年齢を組み合わせた情報は個人特定に繋がりやすく、78 人において一意に個人を特定することが可能であった（表 7）。どちらかの年齢が比較的若いもしくは高齢である場合、または死亡までの年数が短い場合に特定されやすいことも明らかになった。両変数ともに 5 歳刻みにカテゴリ化することで、死亡時年齢が欠損している 1 例を除き、個人の特定可能性について対象可能であった（表 8）。

2-5. ベースライン時年齢と地域

ベースライン時年齢（1 歳刻み）とコホートを組み合わせると、200 人の個人を一意に特定することが可能であった。年齢を 5 歳刻みにしたところ、この人数は 27 人となった。同様に年齢を 10 歳刻みとした場合に、人数が 5 人以下となるパターンを表 9 に示した。ベースライン時年齢が比較的若い場合、またコホート内の死亡が少ない場合に、これら 2 つの情報を組み合わせることで個人の特定可能性が上昇すると考えられた。

2-6. ベースライン時年齢と死亡年

1 歳刻みのベースライン時年齢とコホートコードを組み合わせると、67 人が一意に特定された。年齢を 5 歳刻みにした場合、人数が 5 人以下となる組み合わせを表 10 に示した。ベースライン時年齢と死亡年を組み合わせた場合、年齢カテゴリを荒くすることのみでは個人特定を防ぐことができず、特に追跡早期に死亡した例で特定可能性が高かった。

表 7. 個人が一意に特定できるベースライン時と死亡時年齢の組み合わせ

Index	ベースライン時年齢	死亡時年齢	Index	ベースライン時年齢	死亡時年齢
1	40	46	40	59	66
2	40	47	41	59	68
3	41	44	42	60	69
4	41	48	43	61	61
5	42	46	44	61	70
6	42	48	45	62	71
7	43	51	46	64	64
8	44	46	47	65	65
9	44	53	48	66	66
10	45	47	49	68	68
11	45	48	50	72	82
12	45	49	51	73	73
13	45	50	52	73	82
14	45	52	53	74	83
15	45	53	54	74	84
16	46	47	55	75	86
17	46	50	56	77	77
18	48	51	57	77	87
19	49	50	58	77	88
20	49	51	59	78	88
21	49	55	60	78	90
22	49	57	61	79	89
23	49	58	62	79	90
24	50	56	63	80	90
25	50	58	64	80	91
26	50	59	65	83	92
27	51	52	66	85	85
28	51	55	67	85	95
29	51	59	68	85	96
30	52	53	69	85	97
31	52	60	70	86	95
32	54	55	71	86	96
33	54	63	72	86	97
34	55	55	73	87	87
35	55	64	74	87	97
36	57	57	75	88	96
37	57	66	76	88	98
38	58	58	77	89	98
39	59	60	78	89	99

表 8. 人数が 5 人以下となるベースライン時と死亡時年齢（5 歳刻み）の組み合わせ

ベースライン時年齢カテゴリ	死亡時年齢	人数
40	40	3
70	.	1
75	90	2

表 9. 人数が 5 人以下となるベースライン時年齢（10 歳刻み）と地域の組み合わせ

ベースライン時年齢カテゴリ	コホートコード	人数
40	10	1
40	23	1
40	31	1
40	33	1
40	34	1
50	1	1
50	21	1
50	29.1	1
80	31	1
40	8	2
40	20	2
40	22	2
40	2	3
40	4	3
40	17	3
50	8	3
50	20	3
60	14	3
60	21	3
60	29.1	3
80	21	3
40	5	4
50	33	4
50	34	4
80	29.1	4
50	23	5
70	21	5

表 10. 人数が 5 人以下となるベースライン時年齢（10 歳刻み）と死亡年の組み合わせ

ベースライン時年齢カテゴリ	死亡年	人数
40	2003	1
40	2005	1
45	2000	1
55	2000	1
60	2000	1
70	.	1
80	1999	1
40	2010	2
65	2000	2
70	1999	2
70	2000	2
85	2001	2
40	2004	3
60	2001	3
65	2001	3
50	2002	4
40	2006	5
45	2003	5
50	2003	5
55	2002	5
80	2000	5

2-7. 死亡時年齢と地域

死亡時年齢（1歳刻み）とコホートを組み合わせると、191人の個人が一意に特定された。年齢を5歳刻みにしたところ、この人数は32人となった。年齢を10歳刻みとした場合に、人数が5人以下となるパターンを表11に示した。死亡が少ないコホートにおいて比較的若いまたは高齢の人が死亡した場合で、特定可能性が上昇していた。

2-8. 死亡時年齢と死亡年

死亡時年齢（1歳刻み）とコホートを組み合わせることで、74人が一意に特定された。年齢を5歳刻みにした場合で、人数が5人以下となる組み合わせを表12に示した。特に追跡早期の死亡例で特定性が高く、死亡時年齢だけでなく、死亡年を丸めるなどの対処が必要であると考えられた。

表11. 人数が5人以下となる死亡時年齢（10歳刻み）と地域の組み合わせ

死亡時年齢カテゴリ	コホートコード	人数
40	4	1
40	8	1
40	10	1
40	20	1
40	23	1
40	31	1
40	34	1
50	1	1
50	20	1
50	21	1
50	25	1
50	29.1	1
50	34	1
60	29.1	1
90	21	1
90	31	1
90	33	1
40	13	2
40	17	2
40	30	2
60	21	2
60	27	2
90	29.1	2
40	2	3
40	5	3
50	33	3
80	21	3
90	23	3
40	7	4
50	4	4
50	8	4
50	23	4
50	31	4
50	2	5
60	1	5
70	21	5
90	8	5
90	11	5
90	34	5

表12. 人数が5人以下となる死亡時年齢（5歳刻み）と死亡年の組み合わせ

死亡時年齢カテゴリ	死亡年	人数
40	2003	1
40	2004	1
40	2007	1
45	2000	1
55	2000	1
60	2001	1
70	2000	1
80	1999	1
90	2001	1
95	2004	1
45	2010	2
50	2010	2
65	2001	2
70	1999	2
85	2000	2
95	2006	2
45	2005	3
50	2002	3
65	2000	3
80	2000	3
85	2001	3
95	2007	3
45	2004	4
50	2003	4
60	2002	4
45	2003	5
55	2002	5
95	2008	5

3. 3変数による個人特定可能性の検討

3-1. 性、ベースライン時年齢、死亡時年齢

これら3つの変数を組み合わせる際には、局所的にカテゴリを粗くする等、個別例に関する検討が必要であると考えられた（表13）。

表13. 人数が5人以下となる性、ベースライン時年齢（5歳刻み）、死亡時年齢の組み合わせ

性別	ベースライン時年齢カテゴリ	死亡時年齢カテゴリ	人数
男性	40	40	2
	70		1
	75	90	1
女性	40	40	1
	40	50	3
	45	45	5
	45	55	5
	50	50	5
	50	60	5
	75	90	1

3-2. 性、死亡時年齢、地域

これら3変数の組み合わせ（死亡時年齢は5歳刻み）より、76人について一意の特定が可能であった。死亡時年齢を10歳刻みにしても一意に特定可能であった32例について表14に示した。死亡が少ないコホートについては変数の情報が加わるほど個人が特定される可能性が大きくなるため、情報の開示の可否も含め検討する必要がある。開示を行う場合には近接地域で統合し開示する、地域を数字や記号などを用い秘匿するなどといった対策が必要になると考えられる。

3-3. 性、死亡時年齢、死亡年

死亡時年齢を5歳刻みとしたとき、これら変数の組み合わせより一意に特定できるのは23名であった。10歳刻みとした場合の結果を表15に示した。死亡年を複数年でまとめることで個人特定の可能性を低くすることが可能と考えらえた。

表14. 個人が一意に特定される性、死亡時年齢（10歳刻み）、地域の組み合わせ

性別	死亡時年齢カテゴリ	コホートコード
男性	40	2
	40	4
	40	5
	40	8
	40	20
	40	31
	50	8
	50	25
	50	29.1
	50	34
	60	27
	90	21
	90	23
	90	31
	90	33
女性	40	7
	40	10
	40	23
	40	34
	50	1
	50	4
	50	20
	50	21
	50	22
	50	23
	50	33
	60	1
	60	27
	60	29.1
	70	21
	90	2
	90	34

3-4. 性、地域、死亡年

3-2同様に、コホートは他に2つの変数と組み合わせると個人特定能が高く、この3変数の組み合わせからは43例が一意に特定可能であった。開示にあたっては、死亡年をさらに粗くカテゴリ化するか、もしくは地域に関して3-2で示した措置を施す必要がある。

表 15. 人数が 5 人以下となる性、死亡時年齢（10 歳刻み）、死亡年の組み合わせ

性別	死亡時年齢カテゴリ	死亡年	人数
男性	40	2000	1
	40	2005	1
	50	2000	1
	80	2000	1
	90	2001	1
女性	60	2001	1
	80	1999	1
	80	2001	1
男性	60	2001	2
	70	1999	2
女性	40	2005	2
	40	2006	2
	40	2010	2
	90	2003	2
男性	40	2003	3
	40	2008	3
	60	2000	3
女性	40	2003	3
	40	2008	3
	40	2009	3
	50	2010	3
	70	2000	3
男性	40	2007	4
	40	2009	4
	50	2002	4
	90	2003	4
女性	50	2002	4
	60	2002	4
	80	2000	4
男性	40	2004	5
	70	2000	5
	90	2004	5
女性	40	2007	5
	50	2003	5

3-5. ベースライン時年齢、死亡時年齢、地域

年齢についての 2 変数をそれぞれ 5 歳刻みにすると、95 人を一意に特定することが可能であった。同様に 10 歳刻みにし、一意に個人が特定できる例を表 16 に示した。特定可能例を多く含むコホートがいくつか存在しており、この結果からも 3-2 に示した地域情報に対する措置が必要であると考えられた。

表 16 一意に特定可能なベースライン時と死亡時年齢（10 歳刻み）、地域の組み合わせ

ベースライン時年齢カテゴリ	死亡時年齢カテゴリ	コホートコード
40	40	4
40	40	8
40	40	10
40	40	20
40	40	23
40	40	31
40	40	34
40	50	5
40	50	8
40	50	17
40	50	20
40	50	33
50	50	1
50	50	21
50	50	25
50	50	29.1
50	50	34
50	60	23
60	60	29.1
60	70	21
70	.	33
70	80	21
80	90	21
80	90	31
80	90	33

3-6. ベースライン時年齢、死亡時年齢、死亡年

ベースライン時年齢と死亡時年齢を 5 歳刻みにすることにより一意特定できる個人は 27 人であった。同様に両変数を 10 歳刻みにした場合に、合計人数が 5 人以下であった変数のパターンを表 17 に示した。ベースライン時および死亡時年齢のカテゴリ化に加え、死亡年を 5 年刻みでカテゴリ化することで特定可能性を小さくすることが可能と思われた。

3-7. ベースライン時年齢、地域、死亡年

この変数の組み合わせでは、ベースライン時年齢を 10 歳刻みにした場合でも、149 人が一意に特定可能であった。表 18 に死亡年を 5 年刻みにしても一意に特定可能であった 39 例を示した。コホートの情報は他変数と組み合わせる際に注意が必要であることが考えられる。

表 17. 人数が 5 人以下となるベースライン時と死亡時年齢（10 歳刻み）、死亡年の組み合わせ

ベースライン時年齢カテゴリ	死亡時年齢カテゴリ	死亡年	人数
40	40	2000	1
50	50	2000	1
50	60	2002	1
70	.	.	1
70	90	2009	1
70	90	2010	1
80	80	1999	1
80	90	2001	1
40	40	2010	2
50	60	2003	2
70	70	1999	2
40	40	2005	3
60	60	2000	3
60	60	2001	3
60	70	2001	3
60	70	2002	3
40	50	2004	4
40	50	2005	4
50	50	2010	4
70	80	2001	4
40	40	2004	5
40	50	2006	5
80	80	2000	5

3-8. 死亡年齢、地域、死亡年

コホート変数は、3 変数以上の組み合わせで個人の特定可能性が高くなり、死亡年を複数まとめてカテゴリ化しても個人特定性は高かった。表 19 に死亡時年齢を 10 歳刻み、死亡年を 5 年刻みとして、一意特定が可能であった例を示した。

表 18. 一意特定可能なベースライン時年齢（10 歳刻み）と地域、死亡年の組みあわせ

ベースライン時年齢カテゴリ	コホートコード	死亡年
40	2	2005
40	3	2000
40	4	2000
40	5	2000
40	8	2005
40	8	2010
40	10	2005
40	13	2000
40	20	2000
40	20	2010
40	23	2005
40	28	2000
40	28	2010
40	30	2010
40	31	2005
40	33	2005
40	34	2010
50	1	2000
50	4	2000
50	20	2005
50	21	2005
50	23	2010
50	25	2000
50	29.1	2000
50	33	2010
60	4	2010
60	6	2000
60	21	2000
60	22	2010
70	8	2000
70	20	2000
70	21	2000
70	33	.
70	33	2010
80	23	2010
80	27	1995
80	29.1	2000
80	31	2010
80	33	2000

表 19. 一意特定可能な死亡時年齢（10 歳刻み）と地域、死亡年の組み合わせ

死亡時年齢	コホートコード	死亡年
40	2	2005
40	3	2000
40	4	2000
40	5	2000
40	7	2005
40	8	2005
40	10	2005
40	13	2005
40	13	2010
40	20	2000
40	23	2005
40	31	2005
40	34	2010
50	1	2000
50	4	2000
50	8	2010
50	17	2000
50	17	2010
50	20	2010
50	21	2005
50	25	2000
50	29.1	2000
50	30	2010
50	33	2010
50	34	2005
60	8	2010
60	10	2010
60	21	2000
60	21	2005
60	25	2010
60	29.1	2005
60	33	2000
70	4	2010
70	8	2000
70	21	2000
80	20	2000
80	27	1995
80	33	2000
80	33	2010
90	2	2000
90	5	2000
90	6	2000
90	10	2000
90	17	2000
90	21	2005
90	23	2010
90	31	2010
90	33	2005
90	34	2010

4. 4 変数による個人特定可能性の検討

前節での 3 変数の検討結果より、コホート変数を他の変数と組み合わせて用いると特定可能性が高くなることが明らかになった。さらに変数を 1 つ加える検討として、4-1 では地域情報を含まない組み合わせにおいて、各変数においてカテゴリ化をどの程度粗くすることで個人の特定を防ぐことができるかを検討する。4-2~4-5 では、コホート変数に 3-2 のような処理を行わない場合に、他の変数のカテゴリ化をどの程度粗くすることで個人の特定可能性が変化するかを検討する。

4-1. 性、ベースライン時年齢、死亡時年齢、死亡年

2 つの年齢の変数を 10 歳刻みとし、4 つの変数の組み合わせを検討した場合に、人数が 2 人以下となる組み合わせパターンを表 20 に示した。さらに、死亡年を 5 年刻みのカテゴリ化の処理を行い組み合わせた場合を表 21 に示した。この条件では、人数が 5 人以下となるパターンは 11 個、うち一意に特定される人は 5 パターンであった。

4-2. 地域、性、ベースライン時年齢、死亡時年齢

両年齢変数を 5 歳刻み、10 歳刻みにした場合、一意に特定できる人数は 226 人、58 人であった。

表 20. 人数が 2 人以下となる性、ベースライン時、死亡時年齢（10 歳刻み）、死亡年の組み合わせ

性別	ベースライン時年齢カテゴリ	死亡時年齢カテゴリ	死亡年	人数
男性	40	40	2000	1
	40	40	2005	1
	40	50	2006	1
	50	50	2000	1
	50	60	2002	1
	50	60	2003	1
	60	70	2002	1
	70	.	.	1
	70	90	2010	1
	80	80	2000	1
	80	90	2001	1
女性	40	50	2004	1
	40	50	2005	1
	50	60	2003	1
	60	60	2001	1
	70	90	2009	1
	80	80	1999	1
	80	80	2001	1
男性	60	60	2001	2
	70	70	1999	2
	70	80	2002	2
女性	40	40	2005	2
	40	40	2006	2
	40	40	2010	2
	50	50	2009	2
	50	60	2005	2
	60	70	2002	2
	80	90	2003	2

表 21. 人数が 5 人以下の性、ベースライン時、死亡時年齢（10 歳刻み）、死亡年（5 歳刻み）の組みあわせ

性別	ベースライン時年齢カテゴリ	死亡時年齢カテゴリ	死亡年	人数
男性	70			1
男性	70	90	2010	1
女性	40	50	2000	1
女性	70	90	2005	1
女性	80	80	1995	1
男性	70	70	1995	2
女性	40	40	2010	2
男性	40	50	2000	3
女性	40	40	2000	3
女性	40	50	2010	3
男性	50	50	2010	4

4-3. 地域、性、ベースライン時年齢、死亡年

死亡年が 1 年刻みの場合、ベースライン時年齢が 5 歳、10 歳刻みとした場合に一意に特定される人数は 638 人、318 人であった。また死亡年を 5 年刻みのカテゴリ変数とした場合、年齢が 5 歳、10 歳刻みのときに一意に特定される人数は 176 人、90 人であった。

4-4. 地域、性、死亡時年齢、死亡年

死亡年が 1 年刻み、死亡時年齢が 5 歳、10 歳刻みの場合に一意に特定される人数は 682 人、366 人であった。また死亡年を 5 年刻みのカテゴリ変数、年齢が 5 歳、10 歳刻みのときに一意に特定される人数は 214 人、110 人であった。

4-5. 地域、ベースライン時年齢、死亡時年齢、死亡年

死亡年が 1 年刻みの場合、2 つの年齢変数を 5 歳、10 歳刻みとした場合に一意に特定される人数は 614 人、309 人であった。また死亡年を 5 年刻みのカテゴリ変数とした場合、両年齢変数が 5 歳、10 歳刻みのときに一意に特定される人数は 228 人、90 人であった。

5. 5 変数による個人特定可能性の検討

年齢変数を 10 歳刻み、死亡年を 5 年刻みのカテゴリ変数とし、今回検討した全 5 変数を組み合わせた際に、一意に特定された人は 186 人であった（表 22, 23）。これには、地域の情報が大きく影響しており、地域情報を公開する際には工夫が必要である。

表 22 一意に特定される 5 変数の組み合わせ（男性）

コホートコード	ベースライン時年齢カテゴリ	死亡時年齢カテゴリ	死亡年
1	60	70	2000
1	60	70	2005
2	40	40	2005
2	50	60	2005
2	50	60	2010
2	70	70	2010
2	80	80	2010
2	80	90	2000
3	40	40	2000
4	40	40	2000
4	40	50	2005
4	50	50	2000
4	50	50	2005
4	60	70	2010
4	70	80	2000
5	40	40	2000
5	40	50	2005
5	50	60	2000
5	60	70	2000
5	70	80	2000
6	60	70	2000
6	80	90	2000
7	40	40	2005
8	40	40	2005
8	50	50	2005
8	60	70	2010
10	60	70	2000
10	60	70	2010
10	80	80	2000
13	40	50	2000
13	50	50	2000
13	50	60	2000
13	50	60	2005
13	70	80	2000
14	40	50	2000
14	50	50	2010
17	50	50	2010
17	50	60	2000
17	70	70	2010
17	80	90	2000
20	40	40	2000
20	50	60	2005
20	50	60	2010
20	60	60	2000
20	70	80	2000
20	80	80	2010
20	80	90	2005

コホートコード	ベースライン時年齢カテゴリ	死亡時年齢カテゴリ	死亡年
21	60	60	2000
21	60	60	2005
21	60	70	2005
21	70	70	2000
21	70	80	2005
21	80	90	2005
22	60	70	2000
22	70	70	2010
22	80	80	2000
22	80	80	2010
22	80	90	2010
23	80	90	2005
25	50	50	2000
25	60	60	2010
25	60	70	2010
25	70	70	2010
25	70	80	2000
27	60	60	2005
27	60	70	2010
27	70	90	2010
28	40	50	2000
28	50	50	2010
28	50	60	2000
28	60	70	2000
29.1	50	50	2000
29.1	70	80	2005
29.1	80	80	2000
29.1	80	80	2005
30	40	50	2010
30	50	50	2000
30	50	50	2005
30	50	60	2000
30	50	60	2010
31	40	40	2005
31	50	60	2005
31	80	90	2010
33	50	50	2005
33	50	50	2010
33	50	60	2005
33	60	70	2000
33	70		
33	70	80	2010
33	80	80	2000
33	80	90	2005
34	50	50	2005
34	60	70	2010
34	80	90	2010

表 23 一意に特定される 5 変数の組み合わせ（女性）

コホートコード	ベースライン時年齢カテゴリ	死亡時年齢カテゴリ	死亡年
1	50	50	2000
1	60	60	2000
1	60	70	2000
1	80	80	2005
2	50	60	2000
2	50	60	2005
2	50	60	2010
2	80	80	2000
2	80	90	2005
3	50	50	2000
4	40	50	2005
4	60	60	2000
4	70	70	2000
4	70	80	2000
4	70	80	2010
5	50	50	2000
5	70	70	2000
5	80	90	2000
6	60	70	2005
7	40	40	2000
8	40	50	2010
8	60	60	2005
8	60	60	2010
8	60	70	2005
8	60	70	2010
8	70	70	2000
8	70	70	2010
8	70	80	2010
10	40	40	2005
10	60	60	2010
10	70	70	2010
10	80	90	2000
13	40	40	2005
13	40	40	2010
13	50	60	2010
13	60	60	2000
13	70	70	2010
13	70	80	2000
13	80	80	2010
14	40	50	2000
17	40	50	2005
17	50	50	2000
17	50	60	2000
17	60	60	2000
17	60	70	2000
17	70	70	2000
17	70	70	2010
17	80	80	2010
17	80	90	2010

コホートコード	ベースライン時年齢カテゴリ	死亡時年齢カテゴリ	死亡年
20	40	50	2010
20	50	60	2010
20	60	60	2000
20	60	60	2005
20	60	70	2010
21	50	50	2005
21	70	70	2005
22	50	50	2005
22	60	70	2010
22	80	90	2010
23	40	40	2005
23	50	50	2005
23	50	60	2010
23	60	60	2010
23	60	70	2010
23	70	70	2000
23	80	90	2005
23	80	90	2010
25	50	60	2005
25	60	70	2000
27	60	60	2005
27	70	70	2010
27	70	90	2005
27	80	80	1995
28	40	50	2010
28	50	50	2005
28	60	60	2010
28	80	90	2000
29.1	60	60	2005
29.1	70	70	2005
29.1	70	80	2000
30	50	50	2000
30	50	60	2000
31	50	60	2005
31	60	60	2010
33	40	50	2005
33	50	60	2005
33	60	60	2000
34	40	40	2010
34	50	60	2005
34	60	70	2010
34	70	70	2005
34	80	90	2005

Ⅱ．JALS 対象地域の死亡数と市町村規模の検討

1．少数セルの発生状況

40 歳以上での検討では、①総死亡では、18 市町村、②CVD 死亡では、29 市町村において、10 未満の少数セルが発生した。

上記の検討を死亡数の多い 60 歳以上の対象に限定して同様に行うと、①総死亡では、2 市町村、②CVD 死亡では、9 市町村において少数セルが発生した。40 歳以上の集団での検討に比べて、10 未満の少数セルの発生は、総死亡、CVD 共に減少した。

2．市町村規模での検討

1）40 歳以上での検討

市町村について、人口規模（-4,999, 5,000-9,999, 10,000-29,999, 30,000-49,999, 50,000-)に分け、少数セルの発生パターンを検討した（表 23)。

a:-4999 (9 市町村)、b:5000-9999 (5 市町村) では、総死亡、CVD 死亡ともに集計区分で少数セルが発生していた。c:10000-29999 (8 市町村) では、CVD の区分で少数セルが発生していたが、総死亡の区分では少数セルが発生したのは 4 市町村であった。d:30,000-49,999 (5 市町村)、e:50,000- (4 市町村) では、CVD 死亡では大部分の市町村で少数セルが発生していたが、総死亡では発生がみられなかった。

2）60 歳以上での検討

1）と同様の検討を死亡数の多い 60 歳以上の対象に限定して行うと、5000 人以上の市町村であれば、総死亡、CVD 死亡ともに少数セルの発生を抑えられる可能性が考えられた。

E. 結論

JALS 対象者のうち死亡例について、データ開示による個人の特定の可能性を検討した。検討した変数のうち、地域の情報が特に個人の特定に繋がりやすいことが明らかとなった。地域に関する情報を公表する場合には、事前に何らかの工夫を施す必要があると考えられた。

また、約 30 万の死亡者を対象に市町村規模を考慮して検討した結果では、少数セルの発生は、40～50 歳台で多くみられていたが、人口が 30,000 人規模以上であれば、総死亡においては少数セルの発生はみられなかった。集計対象を死亡数も多くなる 60 歳以上に限定することにより、人口規模が 5,000 人以下の村レベルを除けば、人口規模によらず総死亡、CVD ともに少数セルの発生数を抑えられる可能性が考えられた。

F.研究発表

1. 論文発表
2. 学会発表

いずれもなし

G. 知的財産権の出願・登録状況

（予定を含む。）

1. 特許取得
2. 実用新案登録
3.その他

いずれもなし

表 23　市町村規模と死亡数の関連

コホート番号	人口規模			40 歳以上での検討		60 歳以上での検討	
				総死亡	CVD 死亡	総死亡	CVD 死亡
	人口カテゴリー	人口	75 歳以上人口割合	① 10 歳階級	② 10 歳階級	① 10 歳階級	② 10 歳階級
27	a: −4,999	1,380	23.0	○	○	○	○
25		2,573	20.4	○	○		○
1		2,995	16.8	○	○	○	○
8		3,150	16.0	○	○		○
28		3,920	15.2	○	○		○
7		4,072	17.8	○	○		○
24		4,629	20.0	○	○		
29		4,683	22.2	○	○		○
10		4,936	15.2	○	○		○
26	b: 5,000−9,999	5,243	16.9	○	○		
13		5,706	15.9	○	○		
11		6,789	18.1	○	○		○
14		7,144	13.4	○	○		
20		8,283	11.7	○	○		
9	c: 10,000−29,999	10,868	16.4	○	○		
23		11,132	18.5	○	○		
6		11,489	19.1		○		
22		11,674	13.7	○	○		
19		11,920	22.5	○	○		
12		14,931	18.6		○		
30		17,366	16.6		○		
5		19,584	14.8		○		
4	d: 30,000−49,999	30,988	15.4		○		
15		36,379	17.0		○		
3		38,569	12.7		○		
18		39,981	16.8		○		
31		47,973	8.5				
21	e: 50,000−	52,156	14.9		○		
2		57,912	14.2				
16		65,163	21.2		○		
17		125,515	8.3		○		

○：観察数が 10 未満のセルが発生

厚生労働科学研究費補助金
循環器疾患・糖尿病等生活習慣病対策総合研究事業

追跡終了後コホート研究を用いた共通化データベース基盤整備とその活用に関する研究

平成25年度　総括・分担研究報告書

研究代表者　　玉腰　暁子
平成26（2014）年3月

目　次

平成25年度厚生労働省科学研究費補助金(循環器疾患・糖尿病等生活習慣病対策総合研究事業)
「追跡終了後コホート研究を用いた共通化データベース基盤整備とその活用に関する研究」
分担研究報告書

疫学研究データのアーカイブ化の試み

研究分担者　辻一郎　東北大学大学院医学系研究科
研究分担者　祖父江友孝　大阪大学大学院医学系研究科
研究代表者　玉腰暁子　北海道大学大学院医学研究科
研究協力者　山縣　然太朗　山梨大学大学院医学工学総合研究部

研究要旨

日本疫学会統計利用促進委員会と共同で、疫学研究データをアーカイブ化し、外部に委託して管理・運営する際に定めておくべき内容を、特に共同研究の場合を念頭に、管理者、データ提供者、データ利用者別に整理した。疫学データの個別性、データの持つ背景要因の理解の困難さを考えると、データを全ての人に対しオープンにすることが必ずしも好ましい結果をもたらさない可能性も考えられる。一方で、多くの疫学研究のデータは、多額の費用と多くの人々の協力、長期にわたる研究により得られた貴重な情報である。適切なデータ提供ならびに利用のあり方について、今後も慎重な議論が必要である。

A. 目的

疫学研究で収集された個人単位のデータについて、そのアーカイブ化と利用体制の構築を目指し、試験的な実施を試みる。

B. 方法

日本疫学会統計利用促進委員会(委員長:山縣然太朗)と共同し、外部に委託してデータアーカイブを管理・運営する際に定めておくべき内容を整理する。

C. 結果

既存の疫学研究データをアーカイブ化し、一定のルールの下で公開する目的は、大きく、

1. 公費等を投入し、多くの人々の協力を得て作られた貴重な疫学データの有効活用
2. 若手研究者の育成
3. データの検証

に分けることができる。また、公開の範囲も、無条件に全ての項目を全ての人にオープンにするレベルから、一定の審査等手続きを経て承認された研究者に対して一定の項目を提供するレベルまであり、後者も条件をつけない提供レベルから、共同研究まで、多くの段階が考えられる。そこで、今回は、まずデータ情報を公開することにより、共同研究を行える体制を構築するための要件を検討した。

[管理者]

制度を適切に運営するために、データの管理責任者を置く。さらに、提供されるデータの受け入れ、利用申請の整理と承認等を行うための運営委員会・事務局が必要と考えられる。提供されたデータについては、利用希望者が利用を検討できる範囲での情報をネット等で公開する。提供されるデータ数の増加、利

用希望者の増加に伴い、データ預かり業務や事務作業が増えることが予想される。金銭的な手当についても今後検討が必要と考えられる(例えば、利用者から一定金額を徴収するなど)。

[データ提供者]

データ提供者には、法的、倫理的に問題のないデータを提供するよう、インフォームドコンセントの範囲の確認、ならびに所属機関での倫理審査を求める。また、共同研究利用の申し込みを検討するために必要な情報公開の内容としては、少なくとも、研究の概要、調査方法、母集団、標本数、有効回答率、データ収集時期、項目数、ファイル形式を挙げることができた。また、データの詳細を示す情報(調査票、項目名と変数等)の提供も必要である。カテゴリ共同研究のあり方として、データ提供者の承諾が必要か不要か、結果公表時のオーサーシップに対する希望等に関しても、あらかじめ意思表示を求めることが望ましい。

[データ利用者]

今回は二次利用ではあるが共同研究の位置づけを想定している。利用者は公開されている調査内容に基づき、共同研究の申し入れを行う。データ利用にあたっては、提供者との共同研究利用にかかる契約を交わすとともに、ルールに基づいた誓約書を提出する。なお、利用者は研究成果を論文等で公表しなくてはならない。

これらに関し、運用のためのマニュアルを整備することが必要と考えられた。

D. 考察

疫学研究で得られたデータをアーカイブ化し二次利用体制を整備するため、まず共同研究の場合を例に、検討すべき内容を整理した。

公的研究費により実施されるライフサイエンス分野の研究では、現在、論文発表等で公表された成果に関わる生データの複製物、又は構築した公開用データベースの複製物を、バイオサイエンスデータベースセンターに提供することが求められている。しかし、疫学データの個別性、データの持つ背景要因の理解の困難さを考えると、データを全ての人に対しオープンにすることが必ずしも好ましい結果をもたらさない可能性も考えられる。多くの疫学研究のデータは、多額の費用と多くの人々の協力、長期にわたる研究により得られた貴重な情報である。適切なデータ提供ならびに利用のあり方について、今後も慎重な議論が望まれる。

E. 結論

日本疫学会統計利用促進委員会と共同で、疫学研究データをアーカイブ化し、外部に委託して管理・運営する際に定めておくべき内容を、特に共同研究の場合を念頭に整理した。

F.研究発表

1. 論文発表
2. 学会発表

いずれもなし

G. 知的財産権の出願・登録状況

(予定を含む。)

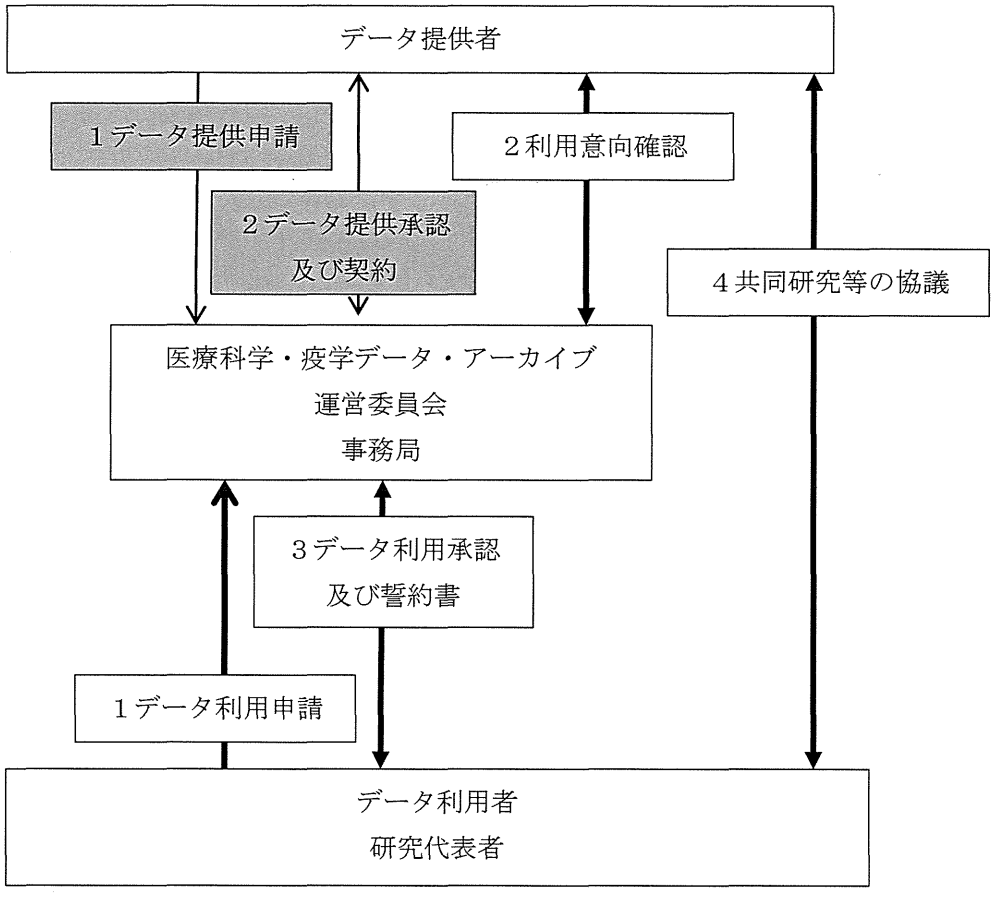
1. 特許取得
2. 実用新案登録
3. その他

いずれもなし

データ提供および利用マニュアル（案）

Ver. 20130820

1．データ提供および利用の流れ



２－1. データ提供手順

1）データ提供申請書および付随する資料を医療科学・疫学データ・アーカイブ事務局（以下事務局）に提出する。

2）データ提供申請を医療科学・疫学データ・アーカイブ運営委員会（以下運営委員会）で審査する。

3）事務局は審査結果をデータ提供者（以下提供者）に通知する。

4）運営委員会が提供を承認した場合は提供者とデータ提供に関する契約を結ぶ。

5）契約後、事務局は提供者から電子データの提供を受け、データに関する付随資料とともに保管する。

２－２. データ提供マニュアル

（1）事前準備

1）申請者は共同研究者に医療科学・疫学データ・アーカイブ（以下データ・アーカイブ）にデータの提供申請をする承認を得る。（注：データ提供申請書に承認を得ていることの一文を入れる）

2）申請者は所属する倫理審査委員会もしくはそれに代わる学会等の倫理審査委員会でデータ・アーカイブにデータを提供する承認を得るなどして、倫理的課題を克服しておく。

（2）提供するもの（データ本体および付随する資料等）

1）データ本体。データはテキストファイル（CSV ファイル形式が望ましい）を基本とする。やむを得ない場合はエクセル、SPSS および SAS のファイル形式を認める。

2）調査票原本もしくはそれに準ずるもの。

3）変数名と調査票の質問との対応表およびデータ入力書式。

4）スコアなどの算出方法。

5）その他データ解析に必要な資料。

（3）データ提供申請

1）データ提供申請者はホームページから WEB 申請を行う。

2）データ提供申請書の書き方は別紙を参照する。

（4）運営委員会の審査結果の受理と不服申し立て

1）申請者は運営委員会の審査結果をメール等で受理し、提供が承認された場合は後述の提供手順に従ってデータおよびそれに付随するものを事務局に提出する。

2）運営委員会は審査結果として、「承認」「条件付き承認」「不承認」のいずれかを申請者に通知する。不承認の場合はその理由を明記する。

3）申請者は審査結果に不服がある場合は、結果通知から 30 日以内に、文書にて事務局に