照合 Record Linkage

照合の問題を考える前に、重複届出、重複腫瘍(多重がん)、及び重複登録の、基礎概念を理解しておくことが重要である。

1. 複数の報告

(1) 重複届出

一人のがん患者における一つの腫瘍に関して登録室が受け取った複数の報告を指す。もしもある患者がある病院でがんと診断され、治療のために別の病院に紹介される場合、両病院とも、その患者をがん登録に届け出ることがよくある。登録室は、これらの報告を重複届出と認識し、集計前に情報を集約する必要がある。

(2) 重複腫瘍(多重がん)

がん登録は、がん患者数というよりも、原発性悪性腫瘍の数を数えるものである。従って、一人の患者が複数の原発性悪性腫瘍に罹患した場合、その各々について独立して登録しなければならない。地域がん登録中央登録室は、原発性悪性腫瘍の過剰及び過小登録を避けるために、多重がんの定義を十分に理解して作業する必要がある。

(3) 重複登録

重複登録は、がん登録が一人の同じ患者を別人として登録したり、一つの同じ腫瘍を誤って重複腫瘍として登録することによって生じる。

2. 照合とは

地域がん登録における照合の目的は、同一人に属する各記録を集めて、その記録が、既に登録されている腫瘍に関するものなのか(重複届出)、あるいは新発生の原発腫瘍に関するものなのかを決定することである。

地域がん登録に集まる情報の例とその登録の流れを図 1 に示した。患者が未登録と判明すれば、新規患者登録が行われ、新たに個人識別番号が付与される。既登録であり、もし記録が、既登録の腫瘍と異なった原発腫瘍(多重がん)に関するものであれば、その腫瘍の情報を別のレコードとして追加登録する。また、既登録であり、腫瘍も既に登録されている場合でも(重複届出)、新たな記録が追加情報(住所や姓の変更、詳細部位や組織型の詳細など)を含んでいる可能性があるので追加登録する。登録室では、患者を管理する個人識別番号と、登録室に集められた個々の記録を区別するシリアル番号の二種類で記録を管理する必要がある。

3. 照合の指標

世界には、社会保険番号等の個人同定番号が国民一人一人に付与されているところがあるが、そのような番号を持たない国において患者が既登録かどうかを確認する基本は、新規患者の姓名と登録室が作成管理する個人識別指標データベースとの照合である。しかし、通常、姓名のみでは不十分である。非常にありふれた姓名、記載間違い、婚姻、離婚、養子による改姓、中国系や韓国系の本名と日本名、及び本人の希望による通称などがその原因である。姓名に加えて、生年月日を照合に用いると識別力が増加する。その他、性別、住所、死亡年月日などが、照合の精度を上げるために用いられる。仮にこれらの指標が全て合致していない

場合でも、記載間違い、婚姻状況、転居等を考慮した結果、同一人物と識別できる場合がある。そのため、これらの照合指標は履歴で管理され、照合の際に常に同一人物の候補として挙がることが望ましい。また、日本人の姓名の場合、同音異字の漢字、旧字、略字、ひら仮名、カタ仮名などの組み合わせで、戸籍上の表記と異なる記載がされている場合が日常的に存在する。例えば、廣島・広島、久子・ひさ子・ヒサ子は、照合の際に同一人物の候補として挙がることが望ましい。

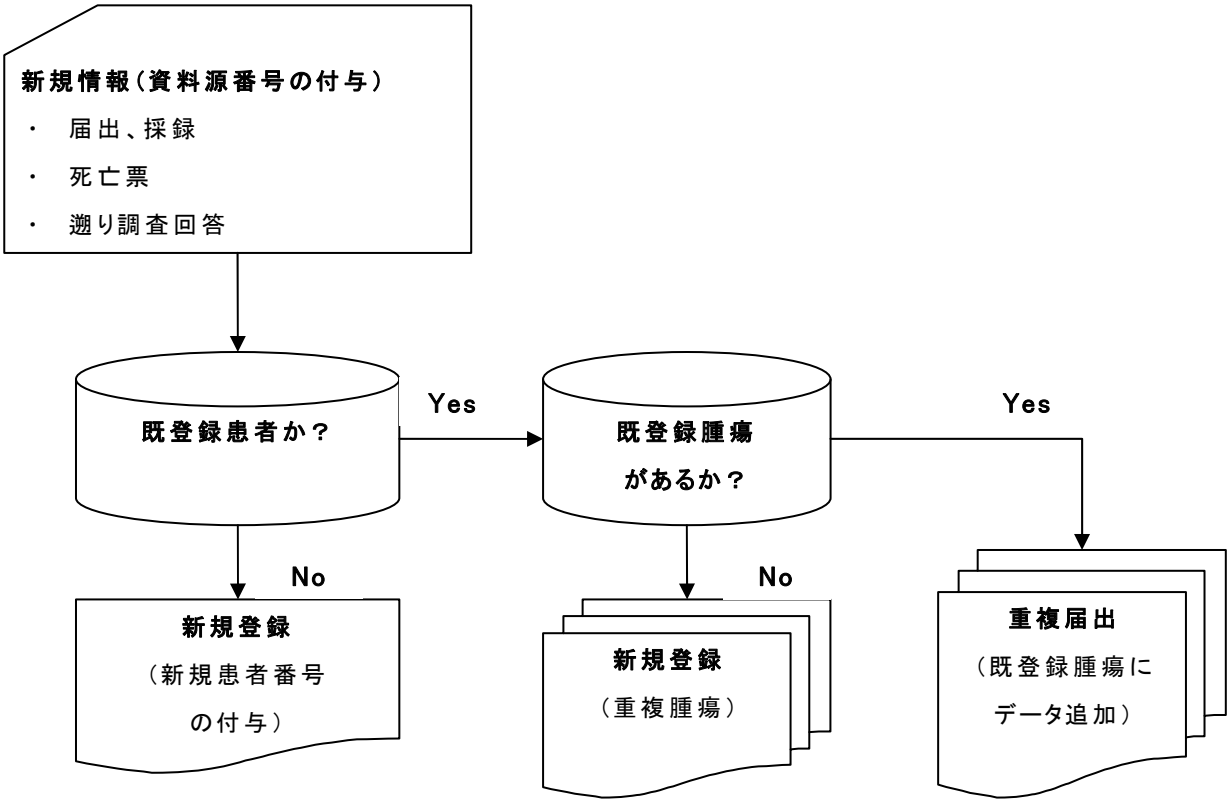

図 1 地域がん登録室における情報の照合の基本的手順

4. 照合の実際

基本的な照合作業は、①個人識別指標を、新しい情報と既登録情報で比較し、同一人物の可能性のある組み合わせを作る作業（同一人物候補リストの作成）、②①で作成された同一人物候補リストについて、他の指標を参考にしながら、同一人物か別人かを判定する作業（判定作業）、の二段階から成る。

従来、照合は、台帳という紙の個人識別指標データベースを用いて、手作業で行われてきたが、今日のがん罹患数とコンピュータデータベースシステムの導入状況を考えると、全て手作業の照合は現実的ではない。ここでは、コンピュータデータベースシステムを用いた照合の実際を記述する。

味木らは、コンピュータによる同一人物候補リストの作成について、わが国の地域がん登録で用いられてきた漢字姓名、生年月日、性別、住所らの個人同定指標の組み合わせを分類し、同一人物候補リストに挙がる数とその作成に要する時間を計測し、精度・効率のよい個人識別指標の組み合わせを検討した（表 1）。コンピュータによる漢字処理に制約の多かった時代、個人識別指標としてカナ表記姓名を用いてきた場合もあるが、漢字表記の方が個人を識別する能力が高いため、今日ではこれを用いる。

同一人物候補リストに挙がる数が少なすぎる

と同一人物が別人として処理されている可能性が高くなり、多すぎるとその後の手作業による同一人物判定作業に時間がかかったり、見落とす可能性が高くなる。表から、日本において照合精度がよい(同一人物を別人と判定する可能性が低い)、かつ照合効率がよい(同一人物候補の数が少ない)方法は、C, D, E 方式であると言える。例として、標準データベースシステムと宮城県地域がん登録における照合方式を提示する。

表 1 照合方式による照合の精度と効率の比較

照合方式	同一人物候補リスト作成に要する時間*	同一人物候補ペア出現数	同一人物ペアの見落とし数（%）
A．生年月日が一致、あるいは姓名が一致	2	23,662	4 (0.1)
B．生年月日が一致した者のうち、他の指標の一致状況を	2		
B-1. 他指標として、性, 住所, 姓名の３指標を利用			
B-1-1. ３指標のうち１指標以上一致		14,144	80 (2.7)
B-1-2. ３指標のうち２指標以上一致		3,571	87 (3.0)
B-2 他指標として、性, 住所, 姓, 名の４指標を利用			
B-2-1. ４指標のうち１指標以上一致		14,157	80 (2.7)
B-2-2. ４指標のうち２指標以上一致		3,640	81 (2.8)
B-2-3. ４指標のうち３指標以上一致		3,242	100 (3.4)
C．B方式を姓名（＋他指標）の一致リストで補足	4		
C-1. 他の３指標のうち１指標以上一致		14,589	4 (0.1)
C-2. 他の３指標のうち２指標以上一致		3,680	11 (0.4)
D．複数指標のうち、一致する項目数が一定値以上			
D-1. 生年西暦, 生月, 姓, 名の４指標を利用	4		
D-1. ４指標のうち３指標以上一致		4,942	18 (0.6)
D-2 元号, 年, 月, 日, 性, 住所, 姓, 名の８指標を利用			
D-2-1. ８指標のうち６指標以上一	28	3,825	5 (0.2)
D-2-2. ８指標のうち７指標以上一致	8	3,327	35 (1.2)
E．大阪府がん登録の方式**	4		
E-1. 氏名としてカナコードを使用		3,371	0 (0.0)
E-2. 氏名として姓の第１漢字を使用		3,325	0 (0.0)
E-3. 氏名として姓名の第１＋３漢字を使用		2,957	4 (0.1)
E-4. 姓名の全漢字を使用（伏せ字は同じとみなす）		2,950	7 (0.2)
E-5, 姓名の全漢字を使用（伏せ字は異なるとみなす）		2,950	7 (0.2)

住所：市区町村単位

大阪府がん登録に、平成４年４月～平成５年３月の一年間に届けられた登録票（24,648件、うち同一人物と判定したペア数2,912件）の資料を用いて検討した。

＊ 生年月日一致リストを作成するために必要な時間を１とした場合

** 生年月日、カナコード（氏名第１漢字に対するカナコード)、住所（市区町村+通字丁)、性別、照合用部位の一致状況の組み合わせが一定のパターンを満たす場合

例1　地域がん登録標準データベースシステム

地域がん登録標準データベースシステムでは、表の D-1 の変形を採用している。概略を示すと、①個人識別指標として漢字姓、漢字名、西暦生年月の 3 指標を用いて同一人物候補リストを抽出、②性別、住所、死亡年月日、届出医療機関コード、原発腫瘍部位コード、組織コード等の判定の補助指標を含むリストを作成、③実務者は②で作成されたリストをもとに判定作業を行う、となる。

標準データベースは、記載間違いなどを考慮し、一度同一人物として処理した個人の漢字姓名、性別、生年月日、住所等個人識別指標を、資料源の情報とともに履歴で保存する機能を持つ。同一人物候補リストにはこれらの履歴情報も併記される。

入力時の条件は、漢字姓名、住所などのテキスト型で登録される日本語一般について、JIS 第二水準までの文字を用いて入力すること、カタカナ・数字・記号も含めて全角で入力することである。第三水準以上の文字は●と処理するか、登録室で第二水準までの代替え文字を決めるなどで対応する。住所は、“字”の挿入や番地の入力方法を統一しておくことが望ましい。山形市大字青柳1800の1と山形市青柳1800－1はコンピュータでは別の住所として認識され、住所履歴の一つとして登録されてしまうのを防ぐためである。

上記のような入力条件を前提として入力されたデータについて、基本データベースに対して姓名の検索を行う前に以下の置換処理を行う（表 2）。漢字には、「広・廣」「斉・斎」などのように異字体が存在するものがあるため、これらの漢字を新字・旧字にも変換する仕組みを置換処理という。各端末に持たせた置換用のテキストファイルにより、検索する姓名の各文字を全ての組み合わせで置き換える。また、ひらがなやカタカナが含まれていれば、表 2 に記した方法で組合せを作る。置換変換数（置換の種類）の上限が姓、名、それぞれ 200 個に設定してあるが、ユキ子などのありふれたカタ仮名名の場合、置換変換数が 200 以上になり機械照合処理が中止されることが稀に生じる。このような事例は、入力時に一時的にゆき子や幸子などを充てることで解決する。表 3 に示す4段階の検索条件で一致した候補者を抽出し（同一人物候補リスト）、最終的に補助指標を併記したリストが作成される（図 2）。実務者は、このリストを目視確認して同一人物を判定する。

表 2　異字体置換処理方法

変換条件	変換内容
1文字目がひらがな	全てカタカナに変換
1文字目がカタカナ	全てひらがなに変換
名の三文字目が″こ″・″コ″	″子″に変換
名の三文字目が″子″	一文字目がひらがななら″こ″に、カタカナなら″コ″に変換
名の二文字目が″ネ″・″ね″	″子″に変換
名の二文字目が″子″	一文字目がひらがななら″ね″に、カタカナなら″ネ″に変換

置換例）　ひさ子 → ①ヒサ子 → ②ひさこ → ③ヒサコ
タネ → ①たね → ②タ子 → ③た子
たね → ①タネ →②た子 → ③タ子

表 3　照合処理の検索条件と分類

一致タイプ	一致項目		
	姓	名	生年月
一致タイプ 1（完全一致）	○	○	○
一致タイプ 2（その他の一致）	○	○	
一致タイプ 3（その他の一致）	○		○
一致タイプ 4（その他の一致）		○	○

図 2　同一人物候補リスト。判定処理用リスト。

例2 宮城県地域がん登録における方法

宮城県地域がん登録では、照合の際、生年月日の①元号、②年、③月、④日、⑤性別、⑥診断時住所(市区町村単位)、⑦姓(漢字)、⑧名(漢字)の8項目を指標として用いている。これらの指標に基づいて6項目以上一致する者が同一人物候補群として抽出され、一致項目数が多い候補者から順に表示される(表1中の D-2-1 に類似した方法)。その際、番地までの診断時住所ならびに現住所、最終生存確認日、死亡年月日等も表示されるとともに、これらの指標に関してこれまでに複数の情報がある者は予備欄に入力されたそれらの情報もあわせて表示される。実務者は表示された照合結果に基づき目視確認にて判定作業を行なう。表示された情報で同一人物か否かの判定がつかない場合には、候補者の原発腫瘍部位コード、組織コード、届出医療機関などの腫瘍情報を確認の上最終的な判定を実施する。なお照合に用いる指標のうち、診断時住所は宮城県居住者については総務省が定めた全国地方公共団体コードの市区町村コードを入力し、用いている。また生年月日に関しては本方法では元号による表記を用いているため、明治 45 年と大正元年、大正 15 年と昭和元年のように同一年が2つの元号にまたがる年については元号、年の表記を予め一方に統一しておくことが望ましいと考えられる。