厚生労働科学研究費補助金
（政策科学総合研究事業（統計情報総合研究））
平成 27 年度～平成 28 年度 総合研究報告書（資料 6）

Probabilistic linkage を用いた大規模公的統計データベースの活用に関する研究

研究協力者　福井敬祐　大阪府立成人病センターがん予防情報センター　研究員
研究分担者　近藤尚己　東京大学大学院医学系研究科　准教授

研究要旨

我が国の人口動態特殊報告を用いた職業・産業別死亡情報は、死亡時職業による分類であるため、実際に長く従事した職業や産業による影響を正確に計ることができないことが懸念されてきた。国勢調査における職業・産業情報と人口動態統計を突合することができれば、その問題点の解決につながる。個人識別番号によるリンケージが不可能な大規模統計データ同士を、個人単位でリンケージする手法として、複数の変数同士の一致確率を統計的に算出して行う probabilistic record linkage の理論と実際について整理した。Probabilistic record linkage においては、リンケージを行うデータベース間での共通のマッチング変数が重要となる。しかしながら、本研究で想定している国勢調査および人口動態統計の二次利用データで入手可能な変数では、あまり正確にリンケージできない可能性がある。また、住所等の情報の入り方の問題もあるため、かなりの作業量を要することが予想された。将来的には Probabilistic record linkage に頼る必要がないよう個人識別番号による突合が可能となるような基盤整備が必要であることが示唆された。

A．研究目的

我が国の人口動態特殊報告を用いた職業・産業別死亡情報は、死亡時職業による分類であるため、実際に長く従事した職業や産業による影響を正確に計ることができないことが懸念されてきた。国勢調査における職業・産業情報と人口動態統計を突合することができれば、その問題点の解決につながるが、現状では氏名・生年月日・住所等、個人をつなぐ情報を使うことができないため、完全なリンケージデータの作成は困難である。

二次利用申請により提供可能である限られた個人属性を用いて、国勢調査データと人口動態調査データをリンケージすることが可能であるかどうか、諸外国で活用されているProbabilistic Record Linkage法について紹介し、我が国における活用可能性を検討する。

B．研究方法

Probabilistic Record Linkageを含む一般的なRecord Linkage手法について、その実行方法を調査し、整理した。特に、Probabilistic Record Linkageについては理論的な背景を中心にまとめ，実行可能なソフト

ウェア等についても紹介した。

（倫理面の配慮）

本報告では実際のデータを用いていないため、倫理面での問題は生じない。

C．研究結果

1. Record Linkage の概要

今、2 つのデータセット A、B を考える。データセットを構成する各データはレコードと呼ばれ、各データセットの中からペアとなるレコードを作成することが Record Linkage の目的である。Linkage に使用する各データセットに共通した変数を Matching Variables (MV)と呼び、MV に基づいて Linkage を行う。例えば、Record Linkage 法として最も単純な Deterministic Record Linkage においては MV の値が同じレコードを Linkage する。図 1 はデータセット A とデータセット B を Record Linkage を行う場合の例である。ここでは ID と名前を MV として、レコード *a* とレコード *b* を Linkage している。

一般的な Linkage の手順は研究毎に様々であるが（例えば、[1, 2, 3]など）それらの方法は簡単に以下の手順に包含することができる (図 2)

① **Data cleaning and standardization**

データの整形にあたる部分である。Linkage を行うデータの誤った入力の修正や、データ間で異なった入力値を共通なものへと変換し、標準化する。

② **Blocking**

各データの間での Linkage 作業における比較数を削減するため、データを Block と呼ばれるいくつかのグループに分割する。Linkage は対応する Block 間のみで行われる。例えば、性別という変数を共通に持つデータ A とデータ B をリンケージする場合には、あらかじめ男性の Block、女性の Block というように Blocking を行う。このようにすることで、Linkage の際には男性の Block 同士、女性の Block 同士のみを比較すればよく、計算量を削減することができる。

③ **Linkage**

Blocking によって作成された Block 間で MV を基にして Linkage 作業を行う。Linkage の方法としては Deterministic な方法と Probabilistic な方法がある。

Deterministic 法は MV が一致するレコード同士をペアとして Linkage する方法であり、Probabilistic 法は MV を基にして作った Weight を使って Linkage を行う方法である。Deterministic 法は非常に単純な方法であり、一般的に MV を種別化・順序化し、データセット同士で結合作業をすればよい。しかし、例えば MV がユニークでない値を含む場合や欠損値が存在する場合などには Linkage の精度は大きく低下する。一般的に Record Linkage を行うデータ量は巨大になりがちであり、データの質の担保が困難な場合が多い。そのため、Deterministic 法を用いた Record Linkage では期待した通りの結果が得られないことが多い。

Probabilistic 法は MV を基に作成された重み（Weight）を利用して Linkage を行う。Weight は Link させたレコードが真のペアであるか否かの確率を反映しており、MV の値を Exact に用いないた

め、データの質が低い場合でも使用できる可能性がある.

④ **Clerical review**

Linkage 作業で Linked もしくは Non-Linked に判別されないようなデータを人為的に判別することを指す。 また、Linkage 作業に用いた閾値などのパラメータ設定やソフトウェアの実行が正しいものであったのかを判別することも含む。

⑤ **Evaluating data quality**

省略

2. Probabilistic Record Linkage について

Probabilistic Record Linkage の基本的な考え方は[4]により提案された。今、$\boldsymbol{A}$、$\boldsymbol{B}$ を Linkage 対象のデータ、$a \in \boldsymbol{A}$、$b \in \boldsymbol{B}$を任意のそれぞれのレコードとする。このとき、直積集合$\boldsymbol{A} \times \boldsymbol{B} = \{(a,b)|a \in \boldsymbol{A}、b \in \boldsymbol{B}\}$の２つの部分集合を

$$\boldsymbol{M} = \{(a,b)|a = b、a \in \boldsymbol{A}、b \in \boldsymbol{B}\}、$$
$$\boldsymbol{U} = \{(a,b)|a \neq b、a \in \boldsymbol{A}、b \in \boldsymbol{B}\}、$$

とすれば、$\boldsymbol{M}$は真に Link 関係にあるレコードの集合、$\boldsymbol{U}$は真には Link 関係にないレコードの集合を表す。また、$\boldsymbol{t}(a,b)$をレコードa,bの一致度を測る一致度ベクトルとし、レコードa,bが真に Link 関係にあるとき、一致度ベクトルが$\boldsymbol{t}$となる確率を$m(\boldsymbol{t})$と定義する。すわわち、

$$m(\boldsymbol{t}) = P(\boldsymbol{t}(a,b) = \boldsymbol{t}|(a,b) \in \boldsymbol{M}).$$

同様に、レコードa,bが真に Link 関係にないときに、一致度ベクトルが$\boldsymbol{t}$となる確率を$u(\boldsymbol{t})$で定義する。すなわち、

$$u(\boldsymbol{t}) = P(\boldsymbol{t}(a,b) = \boldsymbol{t}|(a,b) \in \boldsymbol{U}).$$

このとき、[4]は次の Weight を用いて Link か否かを決定する方法を提案している。

$$w(\boldsymbol{t}) = \log\frac{m(\boldsymbol{t})}{u(\boldsymbol{t})}.$$

実際には、一致度ベクトルが取り得る値$\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k$の全てに対して$w(\boldsymbol{t}_1), \ldots, w(\boldsymbol{t}_k)$を計算し、あらかじめ設定した閾値と比較するという方法をとる。上記の Weight の計算には、レコード内の角変数に独立性を仮定し、EM アルゴリズムを利用する方法が提案されている(参照[5])。

3. 応用ソフトウェア等

Probabilistic Record Linkage を行うことができるソフトウェアは有償・無償のもの含めて様々開発がされている。また一般的な統計ソフトウェアのパッケージとして提供されているなど、導入しやすい。例えば、National Program of Cancer Registries (NPCR) によって開発・提供されている無償の Record Linkage ソフトウェア Link Plus (図 3)は GUI ユーザーインターフェイスで直感的な操作が可能であり、比較的簡単に Record Linkage が可能である。他にも Record Linkage を行える GUI ベースのソフトとしては、無償のものでは D-Dupe、DuDe、Merge Tool Box などがある。その他のソフトについては[6]を参照されたい。また、統計解析ソフトウェアの R 言語における RecordLinkage パッケージは多少のプログラミング知識を有するが、多量なデータを自動で Linkage したい場合や Linkage したデータを直接分析する必要がある場合などに有用である。

これらのソフトウェアを用いる上での注意点としては，無償版のソフトウェアのほとんどが日本語に完全に対応しいるわけではないということである。そのため、入力値や MV として日本語が含まれるデータの Linkage の際にはその精度に対して十分な注意が必要である。

4. 実適用について

Probabilistic Record Linkage の基礎理論は上述したとおり、[4]により提案されており、その歴史は長く、海外を中心に活用されている（例えば、[7,8,9]など）。一方で、日本における適用例はまだ少なく、[10]においては、Record Linkage を必要とする分野がそれぞれ領域固有の知識を必要とするために、学術的な一般化が難しかったこと、 我が国では戸籍制度が整備されており、国勢調査等で人物同定の必要性がほとんどなかったという社会背景をその理由として挙げている.

D．考察

個人識別番号によるリンケージが不可能な大規模統計データベースを、個人単位でリンケージする手法として、複数の変数同士の一致確率を統計的に算出して行うProbabilistic Record Linkageの理論と実際について調べた。

Record Linkage法は上記に上げた手順①～⑤の作業を必要とするが、応用ソフトウェアではこれらの作業を支援することができるものがほとんどである。 しかし，Linkageに有効なMVとして用いられがちな氏名や住所等の情報は基本的に日本語での入力が行われるため，ソフトウェアでMVとして正確に作用しない可能性がある.

本研究でLinkageを考える人口動態統計と国勢調査データを使用する場合、現行でLinkageに使用可能な変数は以下の通りである。

人口動態統計

- 事件簿番号
- 性別
- 生年月日
- 死亡年月日
- 死亡した人の住所
  死亡票：市区町村コード
  オンライン報告分：詳細住所

国勢調査

- 性別
- 生年月
- 調査区番号
- 市区町村（町丁目）情報

氏名や個人識別番号などの利用が困難である我が国の現状においては、利用可能な変数は有用なMVとなりにくく、ユニークでない組み合わせがかなり存在するリンケージデータとなる可能性が高い。住所情報に関しても共通のコード化などの工夫が必要となり、かなりの作業量を要することが想定される。

相当な作業量を要する上に、そのリンケージデータの精度が低いことが想定されるため、将来的には、北欧諸国や英国、米国のように、個人識別番号の整備を経て、各種公的統計のリンケージを公的機関が行い、個人識別可能な情報を削除した匿名化データを利用者に提供する仕組みが必要であると考える。

E．結論

Probabilistic Record Linkageの実行においては、現状の日本の国勢調査と人口動態統計データでは有用な共通のマッチング変数が利用可能でないため、精度の低いリンケージデータとなる可能性がある。Probabilistic Record Linkageにおける理論やソフトウェアの整備が進む一方で、得られる結果の整合性を考慮すれば、将来的には各種統計データベース間での共通個人識別番号の整備およびその活用について、検討していく必要がある。

Ｆ．健康危険情報

Ｇ．研究発表

１．論文発表

２．学会発表

Ｈ．知的財産権の出願・登録状況
（予定を含む）

１．特許取得

なし

２．実用新案登録

なし

３．その他

なし

引用文献

[1] Gu, L., Baxter, R., Vickers, D. and Rainsford, C. (2003). Record linkage: Current practice and future directions. *CSIRO Mathematical and Information Sciences Technical Report*, **3**, 83.

[2] Elfeky, M. G., Verykios, V. S. and Elmagarmid, A. K. (2002). TAILOR: A record linkage toolbox. In Data Engineering, 2002. Proceedings. 18th International Conference on (pp. 17-28). IEEE.

[3] Lee, M. L., Ling, T. W. and Low, W. L. (2000). IntelliClean: a knowledge-based intelligent data cleaner. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 290-294). ACM.

[4] Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64(328), 1183-1210.

[5] Bauman, G. J. (2006). Computation of weights for probabilistic record linkage using the EM algorithm.

[6] Christen, P., (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Springer Science and Business Media.

[7] Whop, L. J., Diaz, A., Baade, P., Garvey, G., Cunningham, J., Brotherton, J. M., ... & Moore, S. P. (2016). Using probabilistic record linkage methods to identify Australian Indigenous women on the Queensland Pap Smear Register: the National Indigenous Cervical Screening Project. BMJ open, 6(2).

[8] Kesinger, M. R., Kumar, R. G., Ritter, A. C., Sperry, J. L., & Wagner, A. K. (2016). Probabilistic Matching Approach to Link Deidentified Data from a Trauma Registry and a Traumatic Brain Injury Model System Center. American Journal of Physical Medicine & Rehabilitation.

[9] Adam, M., Kuehni, C. E., Spoerri, A., Schmidlin, K., Gumy-Pause, F., Brazzola, P., ... & Zwahlen, M. (2015). Socioeconomic Status and Childhood Leukemia Incidence in

Switzerland. Frontiers in oncology, 5.

[10] 相澤彰子，高須淳宏，大山敬三，& 安達淳. (2004). 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, (8), 43-51.

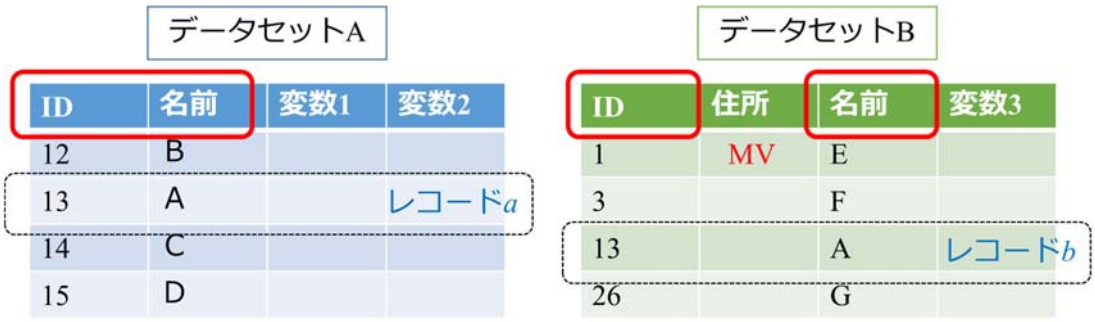

図 1 Record Linkage の例

データセット A 内のレコード a とデータセット B 内のレコード b を ID と名前を基に Linkage

図 2 Record Linkageのフロー図

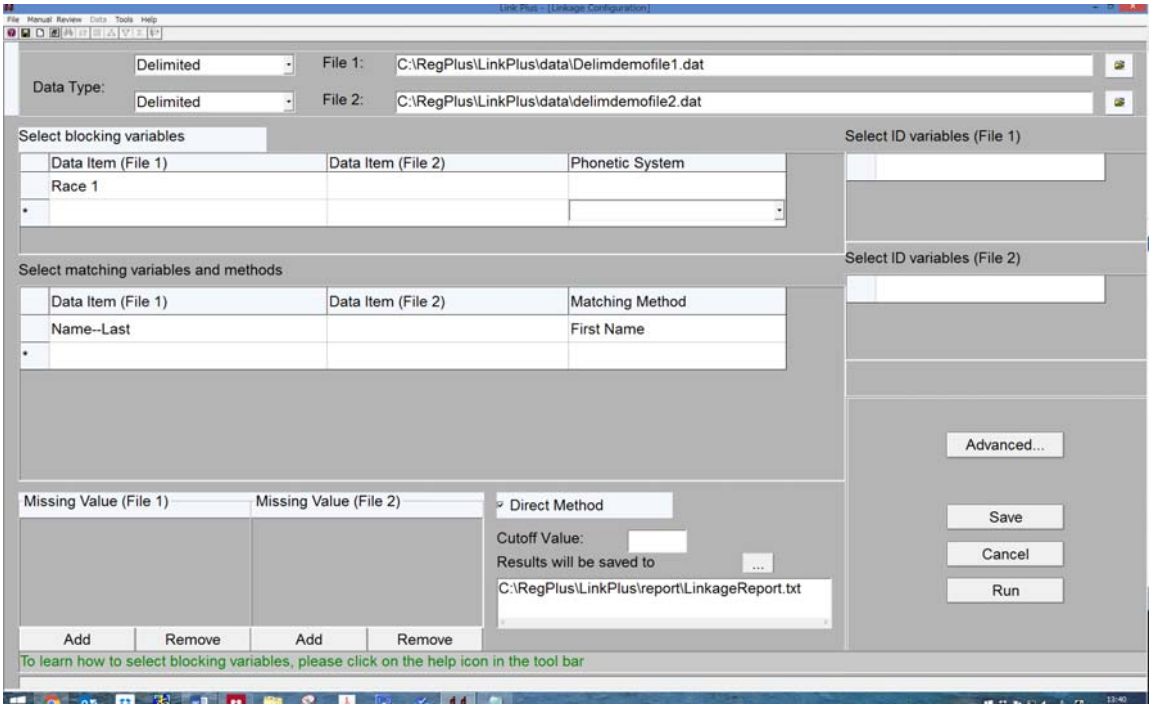

図 3 Link Plus のインターフェイス画面