平成 25 年度 厚生労働科学研究費補助金（成育疾患克服等次世代育成基盤研究事業）
「今後の小児慢性特定疾患治療研究事業のあり方に関する研究」
分担研究報告書

小児慢性特定疾患治療研究事業における Record Linkage 手法の開発と整備

研究分担者　野間 久史（統計数理研究所 データ科学研究系 助教）

研究要旨　本研究では，小児慢性特定疾患治療研究事業で収集されたデータを，外部の公的統計や他研究事業のデータベースと正確にリンクするための標準化された Record Linkage 手法の開発と整備を行う．本年度は，海外の先進的な研究機関で運営されているシステムや，有償のソフトウェアなどの広範な調査を行い，本研究事業で導入するべきシステムについての設計を行うこととした．

結果として，Australian National University が開発した Febrl (Freely Extensible BiomedicalRecord Linkage) が相応しいものと考え，現在，その日本語化についてのプロジェクトを進行中である．新たに開発された日本語化 Febrl は，広く本邦における疫学研究・臨床研究でも利用できるように，汎用性・公共性の高いものとして公開し，本邦における医学研究の発展に資するものとしたいと考えている．

研究協力者:
森 臨太郎（国立成育医療研究センター政策科学研究部長）

A．研究目的

2 次データ（secondary data）を利用した疫学研究の多くでは，複数の異なる情報源からのデータをリンクして，曝露・アウトカムや交絡要因についての情報を揃え，統計的な分析を行うことになるのが一般的である（Olsen, 2008）．小児慢性特定疾患治療研究事業のデータを利用した，疫学研究・臨床研究を実施する際にも，例えば，厚生労働省の人口動態統計などの外部情報を利用することにより，より多くの研究仮説についての研究を行うことができる．これらのプロセスでは，複数のデータベースにある情報を正確にリンクすることが不可欠となってくるが，本研究事業では，米国の社会保障番号（social security number）のような個人を識別する情報が利用できず（本邦における，その他のデータベースも同様である），これらに頼らない，正確な Record Linkage 手法の確立が重要な課題となる．しかしながら，本邦では，諸外国のような公共の Record Linkage システムやソフトウェアはなく，特に，非専門家には，海外の高度なソフトウェアを駆使した解析は，困難でもある．

本研究は，上記のような問題を鑑みて，小児慢性特定疾患治療研究事業における，標準化された Record Linkage 手法の開発と整備を行うものである．その一環として，非専門家にも容易に扱うことのできる，日本語対応したシステム・ソフトウェアを開発する．

B．研究方法

本研究では，海外の先進的な研究機関における Record Linkage システムの運用状況や，情報工学領域における最新の研究動向を調査し，本研究事業における，適切なシステム・ソフトウェアの構築と運用方法，また，実際の研究利用における，標準化された具体的な

手順などの方針を策定する．

また，本研究事業において開発・整備されたRecord Linkageシステムは，より一般的に広く本邦における疫学研究・臨床研究でも利用できるように，汎用性・公共性の高いものとして，Web上に公開し，本邦における医学研究の発展に資することを目的とする．

（倫理面への配慮）

本研究は，方法論やシステム・ソフトウェアの開発が目的であり，実際の患者情報などを利用することはないため，倫理審査は不要と考えられた．

C．研究結果

Record Linkageの統計学的な方法論や計算アルゴリズムについては，古くから研究が行われており，確率的なLinkageの方法も含め，十分に確立された方法論が存在する（詳しくは，Gomatam et al., 2002; Herzog et al., 2007; Liand Shen, 2013などを参照）．本研究事業で運用するシステムでは，これらの方法を十分に標準的な機能として備えたものを構築することが望ましいと考えられる．

海外では，Statistics CanadaのGRLS (Generalized Record Linkage System ; Fair, 2004)やUS Census Bureauのソフトウェアなど，公的な機関が開発したシステムが複数開発されている．また，商用のソフトウェアも多い（Herzog et al., 2007）．これらのソフトウェアでは，一般的に，高額の利用料金がかかる．一方で，RのRecordLinkage（Sariyar and Borg, 2010）のように，フリーのソフトウェアも開発されている．これらを含めれば，海外では，かなりの数のシステム・ソフトウェアが利用可能であり，これまでに，それらの本格的なシステムが一般利用可能な状態となっていない本邦とは格段の差がある．ただし，これらのソフトウェアの多くは，海外で開発されたものであり，日本語で入力されたデータベースのデータ処理に対応していないことなどが難点として挙げられる．また，Rのrecordlinkageのように，特定のプログラム言語に習熟していないと実践での利用が難しいというものもあり，医学・健康科学の分野における統計や計算機に習熟していない研究者やテクニシャンが利用するのは容易ではない．一方で，これらの条件を満たすシステムを新たに構築するためには，膨大なコストと労力が必要となる．

そこで，本研究では，上記のような条件を鑑みて，多くのシステムを精査した結果，Australian National Universityのコンピュータ科学部門のグループが開発したFebrl (Freely Extensible Biomedical Record Linkage; http://datamining.anu.edu.au/)を日本語化して利用することを検討した．Febrlは，比較的新しく開発されたフリーのRecord Linkageのソフトウェアであり，古典的な確率的なLinkageの方法も含めて，最新の機械学習の方法まで，かなり広範な機能が網羅されている（Christen, 2007; 2008）．Febrlは，単にRecord Linkageの技術的なアルゴリズムだけではなく，最も煩雑な，その前段階のデータクリーニングのための機能も充実しており，標準的に使う機能は，概ねそのまま利用することができる．加えて，GUI (Graphical User Interface)によるシステムを備えており，特定のプログラミング言語に習熟しているという必要はなく，Microsoft Excelのような表計算ソフトの上で，データの処理・操作ができる．利用画面のスナップショットを，図1，2に示す．最近でも，システムは定期的に更新されており，追加の機能の充実なども期待することができる．

ただし，問題点として，上記の通り，海外で開発されたソフトウェアとしての例外に漏れず，日本語対応していないことから，本研究事業で即座に利用することはできない．本研究では，Australian National UniversityのFebrlの開発グループの了承を得て，ソフト

ウェアの日本語化を進めており，概ねのところ，来年度にかけて，日本語版 Febrl を完成できる状態に仕上げることを目標としている．

また，日本語化したソフトウェアは，本研究事業のみではなく，より一般的に広く本邦における疫学研究・臨床研究でも利用できるように，汎用性・公共性の高いものとして，フリーソフトウェアとして Web 上に公開し，本邦における医学研究の発展に資するものとしたいと考えている．

D．考察

Record Linkage の方法論の重要性は，古くから認識されていたが，本邦で利用可能な統計においては，米国のような社会保障番号による正確なリンクができないという難点があり，近年でも，薬剤疫学のデータベース研究などでも同様の議論が挙がっている（久保田，2011）．本研究の成果として開発される Record Linkage システムやソフトウェアは，汎用性・公共性の高いものとして，広く我が国における医学研究の発展に資するものとできればと考えている．

一方で，Record Linkage そのものは，対処療法以外の何物でもなく，科学的研究の妥当性を保証するためには，個人を識別する正確な ID などを，省庁・研究事業を問わずに導入するなど，抜本的な改革が要求されるところである．

引用文献

Christen, P. (2007). Febrl—Freely Extensible Biomedical Record Linkage (User Manual; ver. 0.4.01). Department of Computer Science, The Australian NationalUniversity.

Christen, P. (2008). Febrl—Freely Extensible Biomedical Record Linkage. Proceedings of the Australian Workshop on Health Data ana Knowledge Management, Wollongong.

Fair, M. (2004). Generalized record linkage system — Statistics Canada's record linkage software. Austrian Journal of Statistics 33: 37-53.

Gomatam, S., Carter, R., Ariet, M., and Mitchell, G. (2002). An empirical comparison of record linkage procedures. Statistics in Medicine 21: 1485-1496.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007). Data Quality and Record Linkage Techniques. New York, Springer.

久保田潔．(2011)．アジアのデータベースとレコード・リンケージ．薬剤疫学 16: 27-35.

Li, X., and Shen, C. (2013). Linkage of patient records from disparate sources. Statistical Methods in Medical Research 22: 31-38.

Olsen, J. (2008). Using secondary data. In Modern Epidemiology (3rd edn.), Rothman, K. J., Greenland, S., and Lash, T. L., eds. pp. 481-491. Philadelphia, Lippincott Williams & Wilkins.

Sariyar, M. and Borg, A. (2010). The RecordLinkage package: Detecting errors in data. The R Journal 2: 61-67.

E．研究危険情報

なし

F．研究発表

なし

G．知的財産権の出願・登録状況

なし

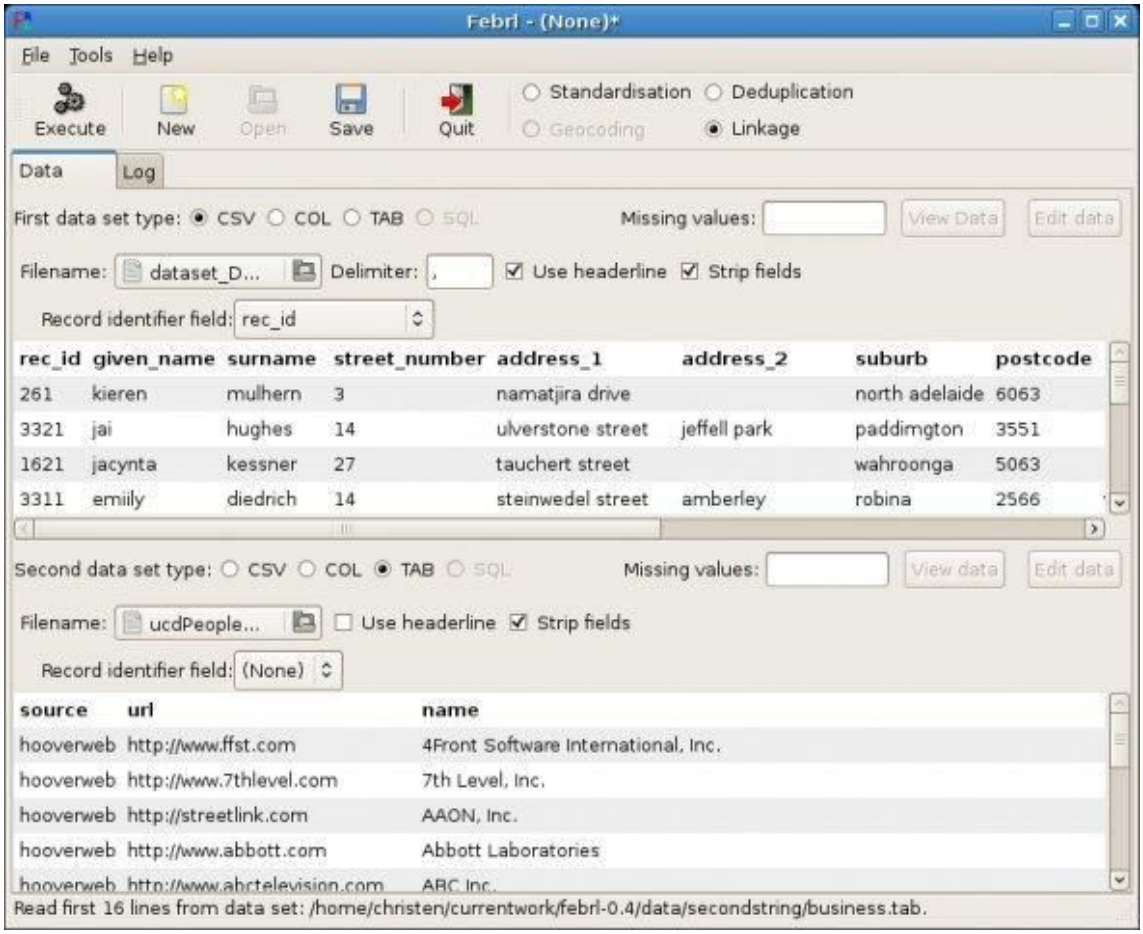

図1　Record Linkage ソフトウェア Febrl のスナップショット．複数のデータベースのデータを，簡単な手順で連結することができる．

（http://sourceforge.net/projects/febrl/）

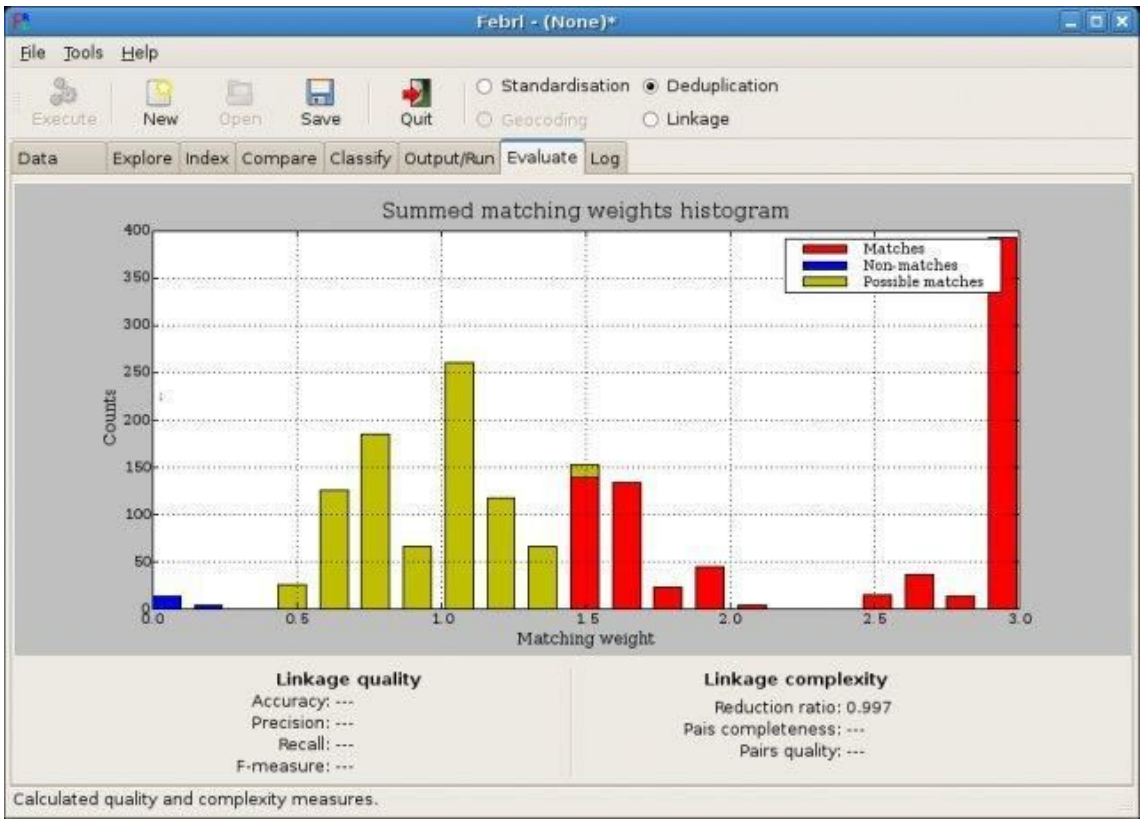

図2　Record Linkage ソフトウェア Febrl のスナップショット．高度な Linkage のアルゴリズムも，簡単な操作で扱うことができる．

（http://sourceforge.net/projects/febrl/）