平成 26 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（健やか次世代育成総合研究事業））
「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 システムの開発と整備：Febrl (Freely Extensible Biomedical Record Linkage) 日本語版の開発

研究分担者: 野間 久史（統計数理研究所データ科学研究系計量科学グループ 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高機能の Record Linkage システムの開発と整備を行う。本年度は，Australian National University によって開発された Febrl (Freely Extensible Biomedical Record Linkage) を日本語対応化し、高度なプログラミング技術がなくても利用することができる、生物・医学領域の研究者にも扱いやすい Graphical User Interface を備えた高機能な Record Linkage システムの基盤整備を行った。また、開発した日本語対応版 Febrl は、本研究の成果物として、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、小児慢性特定疾病情報センターのホームページにフリーソフトウェアとして公開し、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。

A. 研究目的

小児慢性特定疾患治療研究事業のように、疾患登録などの 2 次データ(secondary data)を利用した疫学研究の多くはそれだけでは、十分な治療効果・曝露効果の評価を行うことが困難であり、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 システムの開発と整備を行うものである。その主たる目的は、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を開発し、小児慢性特定疾患治療研究事業のデータを有効に活用した疫学研究・臨床研究のエビデンスを生み出す基盤整備を行うことである。

B. 研究方法

昨年度の研究により、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の調査を行った。これにより、Australian National University のコンピュータ科学部門によって開発された Febrl (Freely Extensible Biomedical Record Linkage) が、本研究の目的に合致した相応しいソフトウェアであると判断し、Python のソースプログラムを改変するなどして、その日本語対応化を行った。

また、開発された日本語対応 Febrl は、本研究事業の成果物として、広く一般の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・公共性の高いパッケージとして、小児慢性特定疾病情報センターのホームページにフリーソフトウェアとして公開する。

（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. 研究結果と考察

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、Christen, 2012; Gomatam et al., 2002; Herzog et al., 2007; Li and 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た。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く流通しているが（Herzog et al., 2007）、これらのソフトウェアでは、一般的に、高額の利用料金が必要となる。一方で、R のパッケージ Record Linkage（Sariyar and Borg, 2010）などのように、フリーのソフトウェアも開発されており、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る。残念ながら、本邦では、これまでは、このような専門的な機能を備えたソフトウェアが公的機関から公開されていなかったというのが実態である。

しかし、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 Record Linkage 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パッケージもあり、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には敷居が高いというのも実情である。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、 Graphical User Interface (GUI) 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、また、高度な連結アルゴリズムを実行することができる。利用画面のスナップショットを、図 1 に示す。先述の通り、最新の高度な連結アルゴリズム（サポートベクトルマシンなど）も実装されており、これらの高機能な計算モジュールを、GUI の直感的なシステムの上で利用することができる。

本研究では、Australian National University のコンピュータ科学部門の Febrl 開発グループの了承を得て、ソフトウェアの日本語化を行った。日本語対応化した Febrl は、オリジナルの Febrl で利用することができる、すべての機能を保持した上で、日本語入力されたデータベースを扱うことができるように改修されている。データセットの入出力も、Excel のフォーマットなどによって作成できるデータテーブルによって可能であり、高度な計算機技術はなくても、容易に取扱いが可能である。また、ソフトウェアのインストールを行う上では、Python などの複数のバックグラウンドモジュールを導入する必要があるが、これらについても専門的な知識がなくても一括してインストールができるように、パッケージ化したインストールモジュールを作製した。導入レベルでの使用のために、日本語による簡易版マニュアルも作成している。

日本語対応 Febrl 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、小児慢性特定疾病情報センターのホームページにフリーソフトウェアとして公開しており（図 2 にスナップショットを示す）、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。また、以上の研究成果は、第 25 回日本疫学会学術総会において報告を行った (Noma and Christen, 2015)。

D. 結論

疫学研究・臨床研究において、Record Linkage 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情報においては、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同様の議論が繰り返し挙がっている（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、長期的な視野で見れば、これらの公的な統計や疾病登録のデータベースを科学的研究に有効活用するために、個人を識別する統一化された ID などを、省庁間・研究事業間を問わずに導入するなどの抜本的な改革が要求されるところである。

E. 引用文献

1) Christen, P. (2007). Febrl－Freely Extensible Biomedical Record Linkage (User Manual; ver. 0.4.01). Department of Computer Science, The Australian National University.
2) Christen, P. (2008). Febrl—Freely Extensible Biomedical Record Linkage. Proceedings of the Australian Workshop on Health Data ana Knowledge Management, Wollongong.
3) Christen, P. (2012). Data Matching—Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, Berlin.
4) Fair, M. (2004). Generalized record linkage system－Statistics Canada's

record linkage software. Austrian Journal of Statistics 33: 37-53.
5) Gomatam, S., Carter, R., Ariet, M., and Mitchell, G. (2002). An empirical comparison of record linkage procedures. Statistics in Medicine 21: 1485-1496.
6) Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007). Data Quality and Record Linkage Techniques. New York, Springer.
7) 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.
8) Li, X., and Shen, C. (2013). Linkage of patient records from disparate sources. Statistical Methods in Medical Research 22: 31-38.
9) Noma, H., and Christen, P. (2015). Febrl Japanese edition: A freely available record linkage system for medical researchers in Japan. Journal of Epidemiology 25 (Suppl. 1): 133.
10) 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.
11) Sariyar, M. and Borg, A. (2010). The RecordLinkage package: Detecting errors in data. The R Journal 2: 61-67.

F. 健康危険情報

なし

G. 研究発表

なし

H. 知的財産権の出願・登録状況

なし

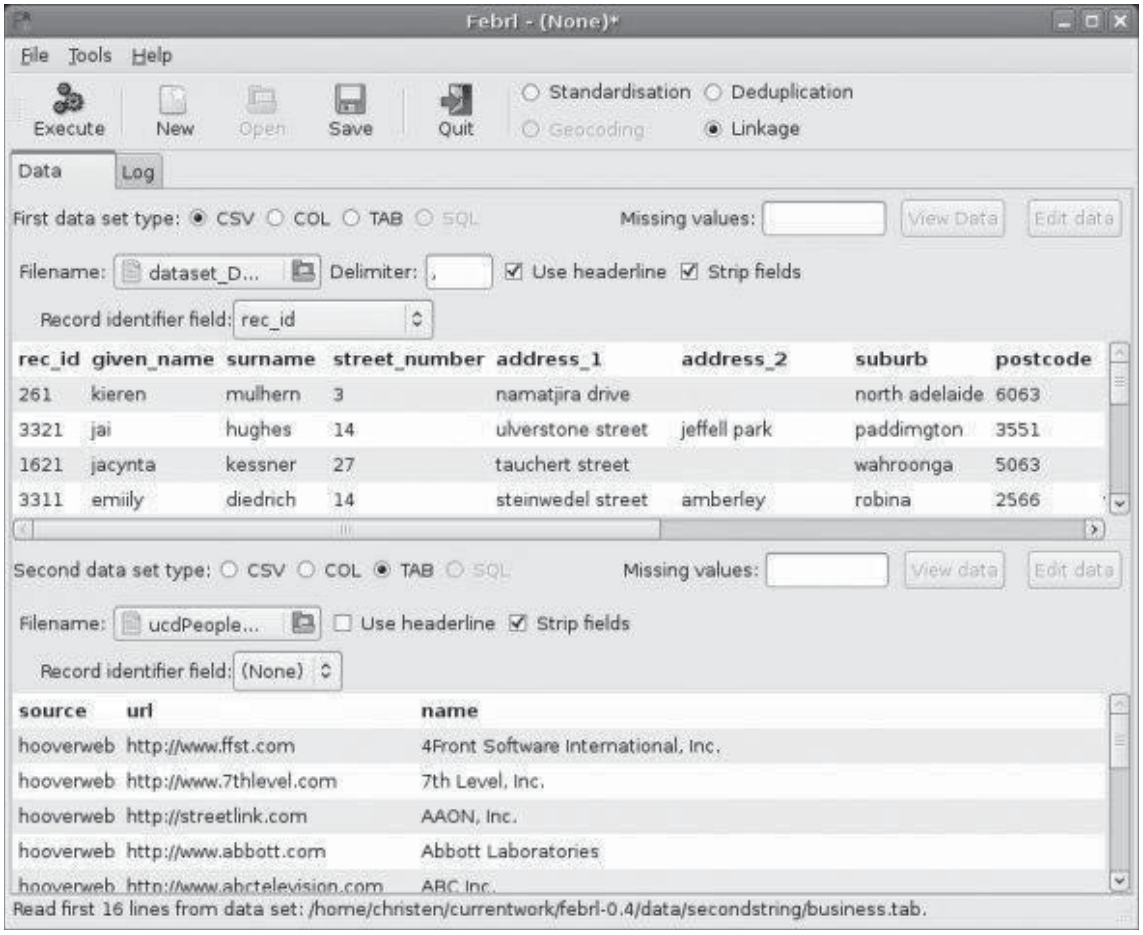

図 1　Record Linkage ソフトウェア Febrl のスナップショット。複数のデータベースのデータを、簡単な手順で連結することができる。
<http://sourceforge.net/projects/febrl/>

図 2　Febrl 日本語版の公開ページのスナップショット。小児慢性特定疾病情報センターのホームページにフリーソフトウェアとして公開しており、広く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。
<http://www.shouman.jp/research/febrl/>