平成 27 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

疫学追跡終了後コホートデータの共通利用（アーカイブ化）の際の死因データ利用に関する検討

研究分担者　大橋 靖雄　中央大学理工学部人間総合理工学科生物統計学
研究協力者　原田亜紀子　東京大学大学院医学系研究科生物統計学

研究要旨

人口動態統計二次利用申請により照合した死因情報を付与した形でのデータアーカイブ化が難しい現状を鑑みて、現制度下での運用案を検討した。一つ目はアーカイブセンターなどで死因情報以外のデータを集約管理し、必要に応じて従来通りの死因照合作業を行い、死因を付加したデータセットを作成し解析を行う方法である。二つ目は、データを保持者のもとに置いたまま、必要と判断した情報だけを選択的に共有させる分散型ネットワークによる方法である。今年度は、後者の分散型ネットワークについて国内外の実例を収集し、疫学研究追跡終了後のコホートデータの共通利用を行う上での課題を検討した。

A. 目的

医療（臨床）データ連携において環境が整備されつつある分散型ネットワークの事例を収集し、疫学研究追跡終了後のコホートデータの共通利用を行う上での課題を検討する。

B. 方法

国内外の医療（臨床）データ連携、疫学共同研究などにおける分散型ネットワークの先行事例を収集した。

C. 結果

1. 国内：医療（臨床）データ連携

1) Standardized Structured Medical record Information eXchange （SS-MIX）

医療機関では電子カルテ・オーダエントリーを中心として、調剤システム、臨床検査システム、画像診断情報システム等、様々な部門システムが稼働し、各々のシステム間で情報がやり取りされている。SS-MIX は、厚生労働省電子的診療情報交換推進事業 (Standardized Structured Medical record Information eXchange)**で策定された『電子的診療情報を他システムとの交換や地域医療連携で利用するために、診療情報を標準的な形式で蓄積・管理するデータとして保存できる領域』の仕様のことである。SS-MIX では、これらの医療情報を「標準化ストレージ」というツールに各情報を標準化した形式で格納・蓄積することにより、複数ベンダー間・複数システム間の相互運用性を高めることを目指している。**標準化ストレージの構造は、図 1 に示す通り、コンピュータの一般的なファイル格納形式と同様に階層化されたフォルダのディレクトリー構造を用いている。電子カルテシステムなどではほとんどの業務で個々の患者を軸として診療情報が格納されているため、標準化ストレージにおいても患者にひも付く各種の情報をフォルダの階層構造にルールを決めて格納している。フォルダの階層構造は、まず各医療施設用のルートフォルダ（医療施設 ID）を置き、配下に患者 ID 先頭 3 文字、患者 ID 4～6 文字、患者 ID、診療日、データ種別フォルダの順に設け、このフォルダの中に HL7 ベースの標準化された各種データファイル群が格納される。**ネットワーク内で「院内リポジトリ」及び「公開リポジトリ」に標準化ストレージを適用し、それ以外の項目について、拡張ストレージを設け統合することで、標準化ストレージを核にし、施設間連携を構

築し、診療情報の研究利用、地域医療連携、災害時連携などを目指している(図 2)。

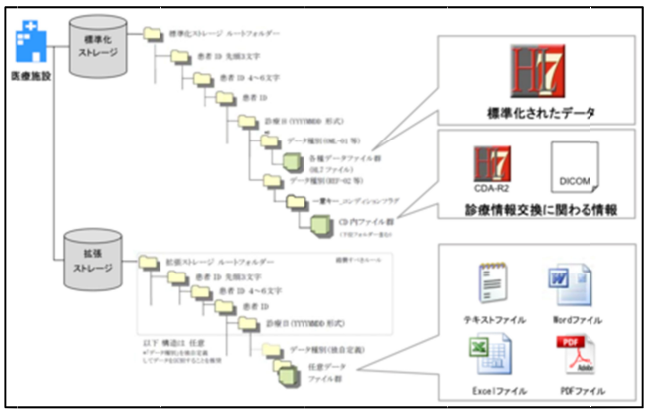

図 1　SS-MIX 情報格納ルール

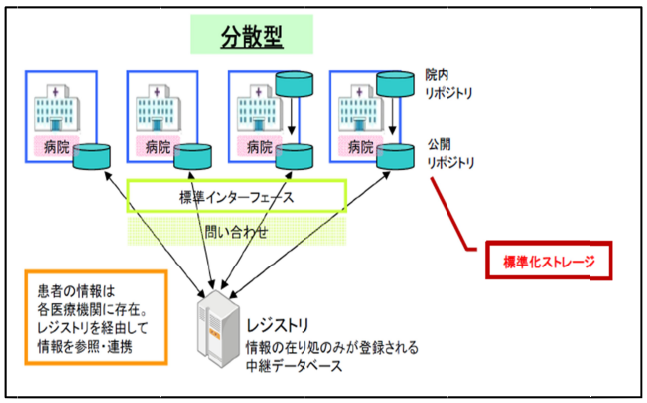

図2　分散型ネットワーク

2. 米国:医療(臨床)データ連携

1) Sentinel Initiative データ交換モデル

米国では 2008 年より、Food and Drug Administration Amendments Act (FDA 改革法)によって FDA 承認医薬品および医療製品の安全性モニタリングのため、既存のデータベースを用いたアクティブ・サーベイランスである Sentinel Initiative が実施されることとなった。Sentinel Initiative ではデータインフラと手法の開発を目的として、Harvard Pilgrim Health Care Institute が中心となって医療データベースを所有する数十の医療機関が協力して Mini-Sentinel が実施され、各データストレージは分散したままで、コモンデータモデルに従って共通プログラムで解析するシステムが構築されている（図 3)。

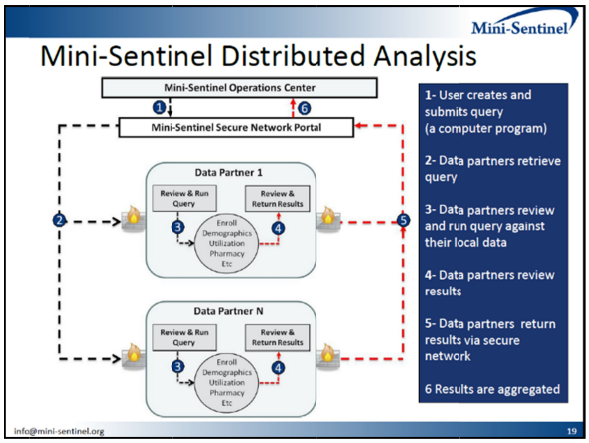

図 3　Mini-Sentinel データ連携モデル

①運用センターが新規 Query をセキュア・ネットワークポータルに発行する
②ローカルサイトが配布された Query を取り込む
③④ローカルサイトにおいて、自施設データに対して Query を実行する
⑤⑥結果を運用センターに戻す
（http://www.mini-sentinel.org より）

2) Health Information Exchange (HIE)

CDX（Crossflo Data Exchange®）ソフトウェアは、Crossflo 社が開発したデータ交換ソフトウェアで、HL7、GJXDM（Global Justice XML Data Model)、NIEM (National Information Exchange Model)、EDXL、CAP、NCPDP など主要医療データシステムおよび国家的なデータ標準、業界標準に準拠したデータ共有方式を実装している。このうち NIEM は、米国政府が後援する、公共および民間部門の組織間での情報共有を促進するためのイニシアチブである。NIEM の当初の目的は法執行、公安、危機管理に置かれていたが、他の分野にも拡大されている。

本システムの医学領域の応用事例としては、モンタナ州および連邦政府の要請を受けて、国立医療情報科学センター の協力のもと、モンタナ州の 4 つの病院の ED（救急科）と州保健福祉局 (Montana Department of Public Health and Human Services）とをデータ接続するモンタナ州医療情報交換システムプロジェクト (Montana Health Information Exchange Pilot Project）があげられる。CDX によるシステムで、ほぼリアルタイムでの医療情報の交換を実現し、異種の医療データソースを迅速および効率的に接続する疾病サーベイランスシステムが構築されている。

3. 国際共同疫学研究等での分散型データ共有

1）ViPAR：Virtual Pooling and Analysis of Research Data (Karter KM et al. *Int.J.Epidemiol.* 2015)

ViPAR は、International Collaboration for Autism Registry において、国際共同研究を実施する6つのサイトデータを統合的に解析することを目的に開発された（Karter KM et al. *Int.J.Epidemiol.* 2015）。多くの医学研究では、倫理的、法的な理由から保有するデータの管理と保持を各サイトにおいて行う必要があるが、本研究では、解析の目的で一時的に「Virtual pooling」サイトに移動することは許可されたことから、一か所にデータを持続的に収集することなしに、物理的に離れたリモートサイトのデータを仮想的(一時的)に集約し分析を可能とするシステムが開発された。

ViPAR のリモートサイトとマスターサーバの関連の関連を図4に示したが、ViPARのマスターサーバは、リモートサイトの PC へ接続されており、研究データはそれぞれリモートサイトで管理されている。利用者は Web 上の analytical portal からアクセスし、データをリモートPCからマスターサーバのRAM上に抽出し、仮想的にデータをプールして解析を行い、終了後は保存することなく消去する。このため外部にデータを収集することなく多施設データを統合した解析が可能となっている。

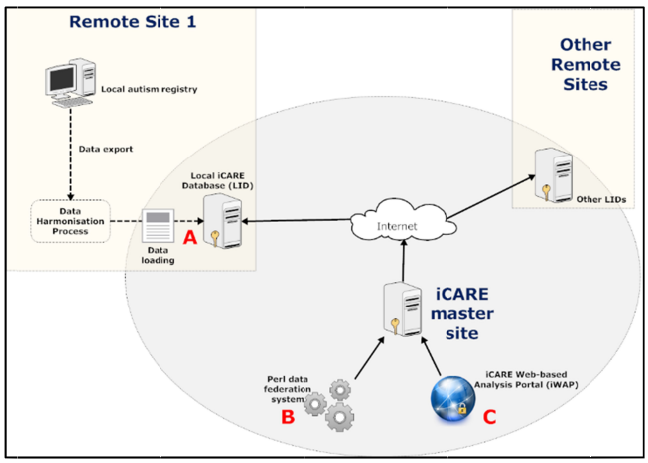

図4 ViPAR のリモートサイトとマスターサーバの関連 (Karter KM et al. *Int.J.Epidemiol.* 2015)

ViPAR の環境のうち、リモートサイトのデータベースは、データ storage として MySQL サーバを、マスターサイトとの通信のため SSH サーバが導入され構成されている（いずれもオープンソース）。実際に解析を行うマスターサーバは、ViPAR daemon と web-based のポータルで構成されている。ViPAR daemon は、ログ出力、統計パッケージのアクセスとコントロール、ローカルサーバから抽出したデータの統合作業を担う。web-based のポータルは、解析、データマネージメントを行うインターフェイスであり、新規の解析スタート画面、解析アウトプット画面、コードのマネージメント画面の三つからなる。ViPAR を用いた具体的な解析の流れは、以下の通りである。

① 解析用インターフェイスを開き変数やサイトを選択
② 統計パッケージの種類（R, SAS, Stata）を選択し、テキスト入力部分に解析のための syntax を入力しサブミットする
③ リモートデータベースにデータのリクエストが送られ、マスターサーバの RAM に読み込まれ、バーチャルプールされる（これらのデータは保存されることはない）。バーチャルプールデータセットは、選択された統計パッケージに読み込まれ、解析が行われる
④ ファイルマネージャー画面で、解析の進行状況、完了状況が確認でき、すべての結果とログがダウンロード可能になっている。

その他の追加の機能としては、よく使用する解析プログラムをアップロードでき、リユースやシェアができ、結果を共有できる機能などが存在する。ViPAR のプログラム、マニュアル等は下記で公開されている
http://bioinformatics.childhealthresearch.org.au/software/vipar/

2) Data SHIELD (Gaye A et al. Int. J. Epidemiol.2014; 43:1929-1944)

DataSHIELD は、研究の管理部門に設置した分析用 PC からローカルサイトにアクセスすることで、ローカルデータを外部に出す（収集する）ことなく統合的に解析するシステムである。既述の ViPAR と類似したシステムであるが、ViPAR が 3 種類の統計パッケージ（R, STATA, SAS）を使用でき、解析を行う際に柔軟に対

応できるのに対し、DataSHIELD は、R 環境で設計されており、使用可能な解析の種類も generalized linear model（GLM）に限られている。ViPAR は、RAM を使用することで解析結果が外部に保存されることはないが、一旦施設外にデータが出ているのに対し、DataSHIELD ではデータがローカルサイトから一度も外に出ることがない点が両者の大きな違いである。

D. 考察

近年整備されつつある分散型ネットワークについて、医療（臨床）データ連携、国際疫学共同研究での先行事例を収集し検討を行ったところ、医療（臨床）と疫学研究の事例では、データ連携(交換)の目的(必要とする背景)、運用方法などが異なっており、本検討課題である疫学追跡終了後コホートデータの共通利用への応用を考える上では、それぞれの特徴を考慮する必要があると思われた。

両者の相違点を整理すると以下のような特徴がみられる。医療（臨床）データ連携と疫学研究データの連携で共通する点は、倫理的規定などからデータ使用の場が制限されており、物理的に施設外に出せないが、連携した利用を積極的に行いたいというニーズがある点である。これに対しては、施設外に出さずにデータ統合と解析を行うさまざまな方法が検討されてきている。医療（臨床）データ連携においては、このようなニーズに加え、データ連携システムの構築を通じて、各種データ形式の標準化を推進していきたいという考えや施設規模の違いや使用しているベンダーの違いを超えた地域医療のデータ連携、災害時のデータバックアップなども目的となっている。さらに、医療（臨床）データ連携では、データの統合方法にも違いがみられ、1）標準化されたデータと統一様式でのデータ保存（階層化、ストレージ作成）を目指す方向（例：SS-MIX）と、2）データの多種多様性を許容したまま、柔軟にデータ連携（交換）を行える交換様式の開発を目指したもの（例：HIE）などが存在している。これらの違いは、連携しようとしているデータ量や連携データを即時的に使用する必要性があるかといったニーズの違いにも依存していると考えられた。

一方、疫学研究でのデータ連携では、医療（臨床）データ連携でみられるような多くの目的を充足する必要がなく、特に即時的なデータ連携、多種多様なデータをそのまま連携したいというニーズは必ずしも高くはない。疫学研究では、研究間で調査項目の共通性・類似性も高いことから、SS-MIX を例に考えるのであれば、共通性の高い項目は標準化ストレージ用の項目、コホート単位でのオリジナルな項目については、拡張ストレージに保存し、ViPAR のようなオープンソースで構築されたシステムを活用して、ローカルサイトにアクセスして共通利用していくような方法も可能なのかもしれない。

逆に医療（臨床）データの連携とは異なる要件としては、大部分の解析をリモートアクセスで行うことを前提にすれば、様々な統計解析を柔軟に行えること、施設外にデータが出るということの定義（法令、指針の解釈）、ローカルサイト（大学、研究所、医療機関を想定）へアクセスを行う上でのセキュリティなどが考えられる。また、人口動態統計の二次利用をローカルサイト単位で各々申請し、死因を付与したデータを統合して用いることや今回紹介した ViPAR での運用のように、施設外部の PC 上の RAM へ一時的な書き出し（保存はされない、電子データの「一時的蓄積」）を行うことの可否など、法令解釈の問題にも留意する必要がある。

一点目の統計解析が柔軟に対応できるかどうかという点については、医療（臨床）データ連携のように、定型化された集計作業（例：Mini sentinel）が中心ではないので、ViPAR のように各種統計パッケージ等と連動可能であるかどうかは重視すべき点である。ただし、ViPAR については、RAM 上の解析で外部にデータ保存が行われないというだけであり、二点目に挙げた施設外にデータが出るということの定義や解釈が利用の際の要件となってくる。三点目のローカルサイト（大学、研究所、医療機関を想定）へアクセスを行う上でのセキュリティの問題については、昨今、施設によっては外部からのアクセスに制限が多くなっている点を考慮すべきである。また、研究によっては、ローカルサイトにセキュリティレベルが低い施設も含む可能性があることなどもこうすべき点である。**四点目の機器利用時・通信過程における一時的なデータ蓄積については、文部科学省文化審議会著作権分科会等でも議論されている。著作権の視点からの議論ではあるが、「一時的固定（複製）」については、次のように整理されている（コンピュータに該当する部分のみ抜粋）。**

ア. 瞬間的・過渡的な蓄積であり「複製」ではないもの
・処理装置(CPU)の読込み
・ビデオ RAM への書込み
イ. 一時的固定（複製）のうち、「複製」と判断すべきものではないもの
・主記憶(RAM)への蓄積
・補助記憶のドライブキャッシュ注釈
・CPU における1次キャッシュ2次キャッシュ
ウ. 一時的固定（複製）のうち、「複製」と判断すべきもの
・主記憶(RAM)への蓄積（常時蓄積）
このような解釈は、技術動向を見極めて判断されるものでもあり難しい課題であるといえる。

ビックデータ、IoT、PHR、地域包括ケアにおける地域医療データ連携などが推進されている現代においては、分散型ネットワークによるデータ連携（交換）は、過去の疫学研究データの統合に限定した話ではない。今後、多施設で行う疫学研究を計画する際にも、一施設にデータを集約しないこのような研究方法の利用は検討していく必要があり、これにより従来の疫学（観察）研究のスキームが、より効率性・生産性が高く、信頼性の高い方法へ向上していく可能性も考えられる。

E. 結論

分散型ネットワークの事例を収集し、疫学追跡終了後コホートデータの共通利用に活用できるかどうか検討を行った。医療（臨床）データ連携と疫学研究の事例では、それぞれ、データ連携(交換)の目的(必要とする背景)、運用方法などが異なっており、疫学研究のデータ連携では、様々な統計解析を柔軟に行えること、ローカルサイト（大学、研究所、医療機関を想定）へアクセスを行う上でのセキュリティ、さらには、人口動態統計の二次利用をローカルサイト単位で各々申請し、死因を付与したデータを統合して用いることや、施設外部の PC 上の RAM への一時的な書き出し（保存はされない、電子データの「一時的蓄積」）の可否など、法令解釈の問題にも留意する必要があると思われた。

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