
一般口演 | データベース・データウェアハウス

一般口演14

データベース・データウェアハウス

2019年11月23日(土) 09:00 ～ 11:00 J会場 (国際展示場 展示ホール8・特設会場2)

[3-J-1-02] 多施設での横断的患者データ検索システムの構築

○松村 泰志¹、松永 隆²、真鍋 史朗¹、村田 泰三³、山本 征司³、小西 正三³、武田 理宏¹（1. 大阪大学大学院医学系研究科 医療情報学, 2. 大阪府立国際がんセンター 医療情報部, 3. 大阪大学医学部附属病院 医療情報部）

キーワード：data warehouse, multicenter study, electronic medical record

大阪大学医学部附属病院では、関連病院と臨床研究を支援するために大阪臨床研究ネットワーク（OCR-net）を組織し、ICT基盤システムを整備しつつある。現時点で19病院の電子カルテシステムと阪大病院のデータセンター間をフレッツ VPNワイドで結び、前向き臨床研究を支援するシステムを導入し運用を開始している。一方、ネットワーク内の医療機関で、被験者候補の検索や薬の副作用解析をするために、後ろ向きに横断的検索を可能とするデータベースの整備も求められている。一方、改正個人情報保護法で病歴が要配慮情報となったことから、各医療機関は、それぞれが管理する診療データを外部に出さなくなった。また、多施設データを横断的に検索するためには、共通のコード体系でデータが保存されている必要がある。こうした状況を考慮し、現実的なソリューションを考案し、整備を進めているので報告する。

レセプト、DPCのEFファイル、様式1のデータには、診療の実施データが保存されており、コードはレセ電算コードで共通化されている。そこで、各施設内に共通データウェアハウスを設置し、患者IDを仮名化し、性、生年月日、処方、注射、手術、処置、病名のデータを格納する方式とした。このデータベースに対して、目的データを取得するための検索処理プログラムを記録したファイルをセンターサーバから各施設のデータウェアハウスに配信し、処理を実行する。処理結果データを各施設のスタッフが確認し、送信を許可した場合、これがデータセンターに集まる仕組みとした。この仕組みを使った検索は、研究を目的としたものに限定し、アウト方式とし、倫理審査で承認を得た場合に実行する運用方針とする。今後、このデータウェアハウスに、検体検査結果、画像レポートなどを追加し、内容を充実させること、秘密計算機能を加えることで、個人情報の収集リスクを減じることを計画している。

多施設での横断的患者データ検索システムの構築

松村泰志^{*1}、真鍋史朗^{*1}、村田泰三^{*2}、山本征司^{*2}、小西正三^{*2}、武田理宏^{*1}、松永 隆^{*3}

^{*1} 大阪大学大学院医学系研究科医療情報学、^{*2} 大阪大学医学部附属病院医療情報部、

^{*3} 大阪国際がんセンター医療情報部

Development of Multicenter Cross-sectional Patient Data Searching System

Yasushi Matsumura^{*1}, Shiro Manabe^{*1}, Taizo Murata^{*2}, Masashi Yamamoto^{*2}, Shouzo Konishi^{*2}, Toshihiro Takeda^{*1}, Takashi Matsunaga^{*3}

^{*1} Department of Medical Informatics, Osaka Graduate School of Medicine,

^{*2} Department of Medical Informatics, Osaka University Hospital,

^{*3} Department of Medical Informatics, Osaka International Cancer Institute

Abstract

We propose a multicenter cross-sectional search system model in which common DWH being installed in each hospital to be processed by common processing program distributed from a center and collect the search results. In common DWH, database files must be composed of a common structure with common code system, thus claim data are useful. Basic patient information, disease name, hospitalization information, prescription, injection, surgical treatment, radiation therapy, and other medical practice data are acquired from claim data, DPC format 1 and EF file. Clinical laboratory test data and image report data are acquired from a data warehouse of each hospital. In these data, house code are used for the item. Thus they are converted to standard code and the value are converted to fit the standard unit by conversion table to create common format database table. According to this model, we are developing the system which will be install in Osaka University Hospital and Osaka International Cancer Research Center.

Keywords: data warehouse, multicenter study, real world data

1. はじめに

これまで医療上のエビデンスの創出には、交絡、バイアスを効果的に除くことができる Randomized Control Trial が採用されてきた。しかし、費用がかかること、倫理的なジレンマが起ることがあることなどの問題があることから、Real World Data を使って、統計的手法で交絡、バイアスを除く方法によりエビデンスを得る方法が注目されるようになった。Real World Data が注目される背景には、電子カルテシステムが世界的に普及し、医療データが電子化されて保存されるようになったことがある。しかし、電子カルテデータは、診療を目的に蓄積されたデータであり、これを臨床研究に利用するためには、解決すべき課題がある。臨床研究のためには、指定する疾患患者のデータを集める必要があり、単施設では十分な数が集まらないことから、多施設からデータを収集する必要がある。電子カルテデータを多施設で集めるためには、そのためのネットワークが必要であること、共通するデータ構造、共通するコードでデータを集める必要がある。

大阪大学医学部附属病院では、関連病院と臨床研究を支援するために大阪臨床研究ネットワーク(OCR-net)を組織し、ICT 基盤システムを整備しつつある。現時点で 19 病院の電子カルテシステムと阪大病院のデータセンター間をフレッツVPN ワイドによるセキュアなネットワークで結び、各電子カルテシステムから、直接データを阪大データセンターに送信することが可能となっている。この環境の上に、前向き臨床研究を支援する目的で、電子カルテ上で共通するテンプレートを開き、入力したデータを標準的な電子症例報告書フォームにマッピングしてデータを集める仕組み (Clinical Data Collection System: CDCS) を導入し、運用を開始している¹⁻³⁾。このシステムでは、多施設の同じ診療科の医師が集まり、共

同して特定の疾患患者のデータを集める患者レジストリを実施することを決めて、データ収集の努力をしようとする場合には、効果的に支援できる。しかし、こうした意図がないところでは、全く無力である。Real World Data という場合には、電子カルテを使ってはいくが、前向きにデータを収集する CDCS のようなモデルではなく、後ろ向きのデータ収集を可能とする環境整備に期待が寄せられている。

例えば、臨床試験を実施する前に、被験者候補の数を推定することが求められるが、そのためには、全患者データから被験者候補になり得る患者を検索できる仕組みが必要である。また、薬の副作用解析をする場合には、ある薬を服用している患者を母数として、有害反応を発生した患者数を数えることが基本となる。既に市販されている薬が、実験研究で他の疾患モデルに効果があることが調べられた場合、たまたま、この疾患の患者で対象薬を服用している患者を調べ、その効果を見ることで、治験を実施する前にある程度の当たりをつけることができる。こうした研究を実施するためには、多施設で全患者データを検索対象とするデータベースを作る必要がある。

我々は、OCR-netに参加する病院で、多施設のデータを横断的に検索できるデータベースの構築を目指してシステム構築を開始している。本稿では、このシステムのモデルについて紹介する。

2. システムモデル

改正個人情報保護法で病歴が要配慮情報となったことから、各医療機関は、それぞれが管理する診療データを外部に出すことが難しくなった。従って、多施設のデータを 1 カ所に集めて横断的に検索可能とする方法は、データ収集した後の処理は簡単になるが、これを実現させることは難しくなっ

きている。現状では、DPC データを収集して解析する研究が実施されているが、それ以上のデータ種を集めようとすると、各病院で躊躇されることが予想される。複数のデータを組み合わせることで、個人が特定されてしまう危険性があること、病院の重要な情報資産を自分達のコントロール下に置いておきたいことがその主な理由である。

そこで、我々は、各病院に、それぞれの病院データを蓄積するデータベースを配置し、データベースをネットワークで結ぶ方法でこの問題を解決させようとしている。各病院に配置するデータベースは、同じデータベースマネジメントシステム上で構築し、共通構造のデータベーステーブルとし、データコードも同じ体系のものとする。これを以下、共通データウェアハウス(共通 DWH)と呼ぶこととする。これにより、ある目的の検索集計をする場合、各病院で同じ検索処理プログラムを適用させることができる。検索処理した結果データを、センターに集めて集計をすることで、多施設の全データを対象とした検索結果と同じ結果を得ることができる(図1)。

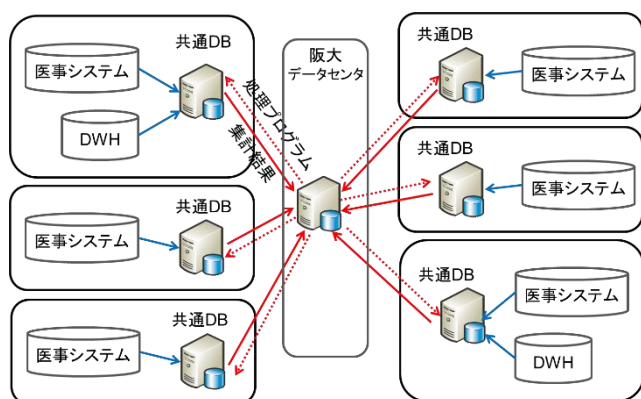


図1 共通 DWH のシステム構成概要

このモデルを実現させるためには、共通コードのデータが必要となる。レセプト、DPC の EF ファイル、様式1のデータには、診療の実施データが保存されており、コードはレセ電算コードで共通化されている。ここから、臨床研究で必要とされるデータを取り出し、共通 DWH に収集する。患者の状態を表すデータは、これらの医事系のデータにはないので、それぞれの病院のデータウェアハウスから抽出して保存する方式とする。当初は、検体検査結果、画像レポートを格納することとし、必要性に応じ、逐次範囲を広げていく予定である。病理レポート、経過記録などを次の目標として考えている。表1に、共通 DWH の当初のデータベースファイル構造を示す。

医事系のデータ以外では、データはハウスコードで管理されている。多施設でデータを集計処理するためには、コードを共通化すること、検体検査結果については、単位をそろえるための変換をすることなどが必要となる。検体検査結果について、全項目に対して対応するためには大変な労力を要するが、実際に各臨床研究で利用される項目は、かなり限定される。そこで、臨床研究で比較利用頻度の多い項目に絞って変換テーブルにデータを登録する方式とした。実際に集計処理をする際には、素データを記録したデータベーステーブルとこの変換テーブルから View フィルを作成することで、共通コードのデータベーステーブルを作成する。このテーブルに対して、処理をすることで、処理を共通化させることができる。この方式により、臨床研究で利用する項目が増えた場

合に、変換テーブルにその項目を追加することで対応ができることとなる。

患者 ID については、処理により1対1に対応する別番号に変換して保存することとした。逆変換処理をすることで、変換後の ID から元の患者 ID に戻せるようにした。これにより、検索した患者について、それぞれの病院で電子カルテから詳細データを確認することができる。

もし、患者が同意しなかった場合には、その患者の ID を登録することで、該当患者のデータを検索対象から外す仕組みを組み込んだ。

我々は、レセプトの病名データと DPC の処方データから治療対象病名を抽出処理するプログラムを開発したが⁴⁾、このシステムも組み込むこととした。

表1 共通 DWH のデータベーステーブル

素データ

①患者基本情報[レセプト]

患者 ID、氏名、氏名カナ、生年月日、性、保険者番号、被保険者証等の記号、レセ年月(=診療年月)

②病名[レセプト]

患者 ID、病名コード、修飾語コード1、修飾語コード2、修飾語コード3、修飾語コード4、病名開始日、未コード病名表記、レセ年月(=診療年月)

③入院[様式1]

患者 ID、入院日、退院日、診療科、身長、体重、退院転帰、レセ年月

④入院 DPC 様式1病名[様式1]

患者 ID、入院日、病名区分※、ICD コード、傷病名コード、修飾語コード、修飾語コード、修飾語コード、修飾語コード、傷病名、レセ年月

⑤処方[EF]

患者 ID、処方日、薬剤コード、薬剤名、使用量、基準単位、実施回数、レセ年月

⑥注射[EF]

患者 ID、実施日、薬剤コード、薬剤名、使用量、基準単位、実施回数、レセ年月

⑦放射線治療[EF]

患者 ID、照射日、診療行為コード、診療行為名称(照射の種類)、解釈番号、レセ年月

⑧手術・処置[EF]

患者 ID、診療行為コード、診療行為名称、解釈番号、実施日、レセ年月

⑨医療行為(処方、注射、放射線治療、手術処置以外)[EF]

患者 ID、診療行為コード、診療行為名称、解釈番号、実施日、レセ年月

⑩検体検査結果[DWH]

患者 ID、性、年齢、検査日時、検査ハウスコード、検査項目名、値、異常フラグ、更新日

⑪画像レポート[DWH]

患者 ID、検査日、画像検査ハウスコード、検査名、モダリティ名、所見内容、サマリ(診断)、更新日

⑫データ利用拒否患者

患者 ID

マスタ

①検査正常値マスタ

行 ID、検査項目ハウスコード、検査項目コード、検査項目

名、性、正常値下限、正常値上限、正常値テキストデータ、オリジナル単位、世代番号、開始日、終了日

②薬剤マスタ

行 ID、医事コード、厚生省 7 桁コード、成分名、成分量

③検査マスタ

行 ID、検査項目ハウスコード、検査項目コード(JLAC10 方法を除く)、検査項目コード(JLAC10 方法コード)、検査項目名、値変換係数、標準単位、世代番号、開始日、終了日

④画像検査

行 ID、画像検査ハウスコード、画像検査標準コード

生成テーブル

①患者基本情報

仮名化患者 ID、生年月日、性

②病名

仮名化患者 ID、病名コード、修飾語コード、病名開始日、未コード病名表記

③処方

仮名化患者 ID、処方日、厚生省7桁コード、薬剤名、成分名、成分量、用量、単位、投与日数

④注射

仮名化患者 ID、注射実施日、厚生省 7 桁コード、薬剤名、成分名、成分量、投与量、単位、回数

⑤放射線治療

仮名化患者 ID、照射日、照射の種類

⑥手術・処置

仮名化患者 ID、診療行為コード

⑦医療行為

仮名化患者 ID、診療行為コード

⑧検体検査結果(指定項目)

仮名化患者 ID、性、検査日、標準検査項目コード、検査項目名、値(変換後)、標準単位、オリジナルの結果値、オリジナル単位、正常値上限、正常値下限、正常値テキスト

⑨画像レポート

仮名化患者 ID、検査日、画像検査標準コード、検査名、所見内容、サマリ(診断)

⑩データ利用拒否患者

仮名化患者 ID

3. 処理操作

データの検索は、阪大病院のデータセンターに設置したセンターシステムから配信する仕組みとした。どこか1カ所で、処理プログラムを走らせ、目的のデータが得られることを確認した後、この処理プログラムを配信用のフォルダーに保存する。この処理プログラムは、各病院からポーリングにより取り込まれ、それぞれの病院の共通 DWH に対して処理した後、結果データを作成する。この結果データを、各病院の担当者が確認し、問題がないと判断した場合に送信ボタンを押す。この操作により、阪大データセンターは、各病院で処理された結果データを受信することになる。研究者が、この結果データを統合処理することで、全体の結果を得ることができる。阪大に設置したセンターサーバの操作は、指定した端末からリモートでもアクセスが可能であり、研究代表者が、阪大の医師でない場合でも対応ができる。

4. 考察

多施設で共通 DWH を立ち上げ、センターから共通の検索処理プログラムを走らせる方法で、データを集計するモデル

について報告した。この方式では、事前に全データを集めることにはならないので、共通 DWH を構築することについては、倫理審査は不要と考える。共通 DWH の利用は、研究目的に限定し、事前に臨床研究倫理審査委員会で承認を得ることを必要条件とする予定である。これまでの方法では、各施設の DWH を利用して目的データを検索し、検索結果を収集して全体をまとめる作業をしていた。共通 DWH を利用する提案の仕組みでデータを収集することは、この個々の処理の手間を省くものの、各ステップで実施していることは従来と変わらない。従って、倫理審査の観点では、受け入れやすく、各病院から理解を得やすいモデルであると期待している。

多施設からデータを集める場合には、匿名化をして誰のデータが分からないようにすることが必要になる。一方、研究の観点からは、ある検索処理をして抽出できた患者が、意図した通りの患者であるかを確認するステップが必要となることが多い。このニーズに応えるためには各施設で実患者 ID と仮名化した番号との対応表を保持しておくことが必要となる。本システムでは、仮名化 ID から実患者 ID に逆変換処理を可能としているため、検索して得た仮名化患者 ID を、逆変換して実患者 ID を取得し、その患者の電子カルテの内容を確認することができる。このことは、匿名化して多施設データを集積する方式に対して利点となる。

多施設でデータを収集する場合、コードが共通化していないことが大きな障壁となる。各病院が標準コードを付けていれば問題とはならないが、対象施設の全てで、こうした標準コードが付けられていることは期待できない。一方、レセプトデータ、DPC データは、レセプトコードでコードが共通化されている。ここから、できるだけ広くデータを取得しておくことは、共通 DWH を構築する上で現実的な戦略と考える。EF ファイルには、診療行為が実施日と伴に記載されているので、レセプトよりも、この目的のためには適している。しかし、例えば処方内容では、EF ファイルでは1日量と処方日数しか分からないのに対し、処方オーダのデータからは、用法のデータも得ることができる。しかし、臨床研究においては、1日量が分かれば十分であり、用法データまでを必要としないものが多い。従って、EF ファイルの処方、注射データで十分な場合が多い。

レセプト、DPC データでは、患者の状態を表すデータがないので、このデータは、DWH から取得する必要がある。しかし、医事系以外のデータは、基本的にハウスコードで管理されているので、これを標準コードに変換する処理が必要となる。検体検査結果データがその例となるが、病院が採用している検査項目数は多く、これら全てに標準コードを付けるのには大変な労力がかかる。そこで、臨床研究で利用する可能性の高いものに絞って標準コードを振り、それ以外の項目は、必要になった際に標準コードを見つけ、登録していく運用が現実的と考える。本システムでは、後から標準コードを追加した場合でも、View ファイルは、その都度生成されるために問題はない。SS-MIX2でデータを集める方式が MID-NET 等で採用されているが、ストレージの場合には、後から標準コードを追加することができないことが問題である。過去データを含めて、標準コードを付け直すためには、一から作り直すことが必要になる。本方式の、標準コードを後から追加できる性質は、大きなメリットである。

電子カルテ上で、どの病院でも構造化されているデータは、検体検査結果ぐらいしかない。他のデータ種については、一部の病院で構造化されていることはあっても、全病院の範囲では、構造化されていることは期待しにくい。一方、データ検索は、非構造化データ(自由文)に対しては、キーワード検索

しかできず、欲しいデータを取得することができない。この問題を解決するためには、自然言語処理により非構造化データの構造化データへ変換して検索する方法が必要である。後ろ向きデータベースを充実させるためには、自然言語処理による非構造化データの構造化データへの変換技術の開発と並行して進める必要がある。

現在では、秘密計算が現実にご利用できるようになってきている。この技術を使うことで、各施設の検索結果を見ずに、集計処理をすることができる。この技術が利用できると、幾つかのメリットがある。一つには、ある条件で検索をした際に、ある施設で一人だけが該当した場合に、個人が識別できたことになり、このデータを施設から外部に出すことができなくなる。秘密計算を使うと、こうした問題が起こらない。ある患者が複数の医療施設にまたがって診療を受けており、これを連結させて処理をしたい場合、個人識別情報を使用する必要があるが、本人の同意が得られていない場合には、その実施は倫理審査で拒絶される可能性が高いが、秘密計算を使って突合し処理をして統計データにまで進める方法では、理屈上は倫理的に許容される可能性がある。次のステップとして、多施設データを秘密計算で処理するシステムの構築を計画している。

共通 DWH は、最初に阪大病院と大阪府立国際がんセンターに導入し、目標としている処理ができるかを確認し、その後、徐々に OCR-net 参加病院に承諾を得た上で展開していく予定である。

9. 結論

共通 DWH を各施設に導入し、多施設データを横断的に検索可能とするシステムのモデルを示した。現在、このモデルに従ってシステムを構築中である。

謝辞

本技術開発には、2018 年 10 月 11 日から参画した SIP AI ホスピタルの研究開発資金を一部充当し、開発を推進している。

参考文献

- 1) Matsumura Y, Hattori A, Manabe S, Takeda T, Takahashi D, Yamamoto Y, Murata T, Mihara N. Interconnection of Electronic Medical Record with Clinical Data Management System by CDISC ODM. Stud Health Technol Inform. 2014;205:868-72.
- 2) Matsumura Y, Hattori A, Manabe S, Tsuda T, Takeda T, Okada K, Murata T, Mihara N. A strategy for reusing the data of electronic medical record systems for clinical research. Stud Health Technol Inform. 2016;228; 297-301 公益社団法人日本医
- 3) Matsumura Y, Hattori A, Manabe S, Takahashi D, Yamamoto Y, Murata T, Nakagawa A, Mihara N, Takeda T. Case report form reporter: A key component for the integration of electronic medical records and the electronic data capture system. Studies in Health Technology and Informatics, 2017: 245:516-520
- 4) Takeda T, Mihara N, Murata T, Shimai Y, Okada K, Manabe S, Matsumura Y. Estimating the ratio of patients with a certain disease between hospitals for the allocation of patients to clinical trials using health insurance claims data in Japan. Stud Health Technol Inform. 2016;228; 537-541