公募企画

公募企画シンポジウム1
レセプト情報等オンサイトリサーチセンターの現況および今後について ー これまでの進捗の報告、および今後の第三者利用についてー

2017年11月21日(火) 08:30 〜 10:00 E会場 (10F 会議室1003)

[2-E-1-PS1-2] レセプト情報等オンサイトリサーチセンターの現況および今後について ーこれまでの進捗の報告、および今後の第三者利用についてー

松居 宏樹[1], 佐藤 大介[2], 大江 和彦[2] （1.東京大学大学院医学系研究科公共健康医学専攻臨床疫学・経済学, 2.東京大学医学部附属病院 企画情報運営部）

東京大学では、平成27年度より NDBオンサイトセンターの運営・利用に向けたシステムの構築を行ってきた。そのなかで、我々はオンサイトセンターのパフォーマンス研究として NDBデータの特性や抽出に必要となる処理時間などを報告してきた。
現在我々は、模擬申し出として厚労省に届け出た個別の研究課題を遂行している。模擬申し出に取り組む過程において、研究者がレセプト情報を利用する際、データハンドリングプロセスの標準化が大きな課題であることがわかってきた。
このプレゼンテーションでは、模擬申し出に取り組む中で構築した、 NDBをはじめとするレセプト情報ハンドリングの標準プログラムに関してその要件・運用上パフォーマンスを報告する。
レセプト情報の研究を行うにあたり、研究に用いる情報を、年齢・性別などの患者背景情報、処方薬・処置などの処方情報、診断名をはじめとする病名情報と定義した。また、研究者がフォーカスする情報の粒度を、患者個人単位・月単位・エピソード単位・日単位と定義した。研究者が、上記の情報を適切な粒度で集計を行うためのコーディングを行うことは困難である。我々は、研究者が以下の情報を指定することで、データの抽出集計プロセスのある程度自動化している。
・特定の条件に合致するレセプトを抽出するクエリ
・必要な処方情報の一覧とカテゴリ
・必要な病名情報の一覧とカテゴリ
自動化のために作成したプログラムは、研究者が設定した条件のレセプトを抽出し、そのレセプトが算定された個人のすべてのレセプトを各集計単位で集計するバッチ処理である。集計する情報は、処方情報と病名情報をカテゴリ単位で集計し、処方情報は処方開始日、病名情報は診断開始日を集計する。
実運用においては、150万人程度の症例規模のデータを集計するのにおおよそ30時間程度の時間を有した。今後、抽出したデータを元に解析を進める。

レセプト情報等オンサイトリサーチセンターの現況および今後について
－ これまでの進捗の報告、および今後の第三者利用について －

松居 宏樹*1、佐藤 大介*2、大江 和彦*2

*1 東京大学大学院公共健康医学専攻臨床疫学・経済学

*2 東京大学大学院公共健康医学専攻医療情報システム学分野

Current status and future prospects of on-site research centers for the Japanese National Insurance Claims Database

Matsui Hiroki[*1], Sato Daisuke[*2], Ohe Kazuhiko[*2]

[*1]Department of Clinical Epidemiology and Health Economics, School of Public Health, The University of Tokyo

[*2]Department of Health Informatics, School of Public Health, The University of Tokyo

Ministry of Health, Labour and Welfare has established on-site research centers, which aim to support health care researchers to perform clinical, epidemiological, health policy and health economics studies using the Japanese national administrative claim database (National insurance claims Data Base, NDB). In this presentation, we will show the overview of the current system of the on-site research centers. We will also introduce our newly developed data handling application and its performance.

Keywords: Administrative claim database、 Data handling、 Epidemiology

1 緒論

厚生労働省は、これまでレセプト情報を研究者、行政機関等へ提供してきたが、データ提供を受けるためには、研究者側が十分なセキュリティ環境を整備する必要がある等環境面でのハードルが高かった。レセプト情報の研究利用を促進するため平成27年度より国内にレセプト情報等オンサイトリサーチセンター（以下オンサイトセンター）が設置された。オンサイトセンターでは、十分なセキュリティ環境が用意され、研究者は医科・歯科・調剤・DPC・特定健診・保健指導の全レセプトデータ（以下、The Japanese National Insurance Claims Database; NDB）が保管されているサーバーに対してアクセスし、レセプト情報等の提供に関する有識者会議にて承認された内容にそって、それらを解析することができるようになりデータの研究利用に伴う環境面でのハードルが低下したと考えられている。

東京大学では、平成27年度より NDB オンサイトセンターの運営・利用に向けたシステムの構築を行ってきた。そのなかで、我々はオンサイトセンターのパフォーマンス研究として NDB データの特性や抽出に必要となる処理時間、オンサイトリサーチセンターで使用可能なデータ解析ツールやその使用感などを報告してきた。

オンサイトリサーチセンターでは「Business Intelligence (BI) ツール」「Oracle R Enterprise (ORE)」「SQL*Plus」を用いて NDB へアクセスすることができる(1)。このうち、BI ツールは一般的な研究者のニーズに対応する機能を有しておらず、研究利用には限界があること、それに対して、SQL を用いたデータ抽出・解析は、研究に必要なデータセットを作成することができるが、レセプトデータの構造や仕様に関する十分な知識と技術が必要であることが明らかとなり、NDB を研究者が利用するうえでの技術面でのハードルは未だに高いといえる。昨年度、医療情報学会において我々は、オンサイトセンターの運用経験を踏まえ、NDB を研究利用する場合にユーザーが有しているべき技術要件を表1のように整理した(1)。

レセプトデータの解析を行うために、一般的な医療系の研究者が、SQL をはじめとする大規模データハンドリング技術を習得し、レセプトの構造をすべて理解することは非効率である。そこで、我々は研究者が研究に用いるデータを抽出・整形する、NDB オンサイトセンター（東京）にて動作するアプリケーションの開発を行った。基本構想は 2015 年の医療情報学会にて報告を行った(2)が、実運用でのパフォーマンスなどは明らかではなかった。このプレゼンテーションでは、上記アプリケーション関してその要件・運用上のパフォーマンスを、利用実績をもとにしたケースレポートとして報告する。

2 目的

このプレゼンテーションでは以下の点について報告する。

1) レセプトデータ抽出・整形アプリケーションの仕様
2) 臨床疫学研究・医療経済学研究での、アプリケーションの使用感（ケースレポート）

3. 方法

3.1 レセプトデータ抽出・整形アプリケーションの仕様

研究者が必要とする情報を抽出するため、以下の3ステップのクエリを作成する。

1) 任意の条件で、レセプト ID 一覧を抽出するクエリ
2) 抽出したレセプト ID を基に、匿名化個人 ID 一覧を作成するクエリ
3) 匿名化個人 ID 一覧から、関連する集計期間内のレセプト情報を抽出するクエリ

1のクエリは、研究者がひな形を参考に、クエリを自ら作成する運用形態をとった。2～3のクエリの作成はアプリケーション側で自動化した。1～3のクエリを実行することで、任意の条件に該当する症例の一連の時系列情報を抽出できる仕様とした。

次に我々は、研究者が利用しやすいデータの粒度として「レセプト単位データ」「患者実施日単位のデータ」「患者エピソード単位のデータ」を想定した仕様を策定した。

レセプト単位の情報

レセプト識別 ID を情報の最小粒度とし，性別，生年月，観察開始月，観察終了月などの患者背景情報に加え、任意条件で指定した当該レセプトに出現した病名情報と診療行為・薬剤などのレセプト実績をカラムとして含む。

患者実施日レベルの情報

患者の外来受診や，入院中各日単位を最小粒度とし，各日のレセプト実施情報をカラムとして含む．各レセプトに日単位情報が含まれる平成 24 年度以降データについて作成した．

患者エピソードレベルの情報

患者の外来受診や入院、調剤薬局での調剤等のエピソードを情報の最小粒度とし，エピソードの区分，医療機関情報，開始日， 終了日，死亡転帰有無，エピソード中の総医療費，併存症情報，病名情報，レセプト実績等をカラムとして含む．エピソード単位の算出には日単位のレセプト実績情報が必須であるため各レセプトに日単位情報が含まれる平成 24 年度以降データについて作成した．

アプリケーショは、任意条件で抽出した症例の一連の時系列情報を上記粒度で整理するためのクエリを出力する仕様とした。

3.2 整形したデータの利用

作成したクエリを実行し作成されたテーブルを、SQL*Plus から研究者が利用できるようにした。SQL*Plus で整形したデータは ORE を用いて解析することができる。

3.3 アプリケーションの使用感

当該システムを用いて、厚生労働省有識者会議に申請した「医療データの統合・解析による将来予測ミクロシミュレータの構築」、「周術期口腔機能管理による術後肺炎発症予防の効果」の模擬申し出に用いるデータを抽出・整形しデータの解析を行った。

それぞれの研究に必要であったデータの形式、対象症例数及び集計期間、抽出処理時間、テーブルのサイズ、一時解析結果の取得までに必要となった期間を報告する。また、それぞれの研究に従事した研究者の有したスキルレベルに関しても報告する。

4 結果

「医療データの統合・解析による将来予測ミクロシミュレータの構築」に必要となった症例は、個人識別 ID ベースで 0.3%ランダムサンプリングされたデータ（約 500 万人）であり集計期間は平成 25 年 4 月～平成 28 年 3 月、レセプト単位の病名情報が必要であった。抽出に必要となった時間はおよそ 18 時間であり、抽出されたテーブルのサイズはおよそ 5.3GB であった。一時解析の結果取得までに必要となった期間はおよそ 2 か月程度であった。データを使用した研究者は過去に SQL・R を使った経験がなく、今回参考資料を利用しながらレセプト単位テーブルから、データの抽出・集計を SQL で行い、その結果をもとに R を使って解析を行った。

「周術期口腔機能管理による術後肺炎発症予防の効果」に必要となった症例は、平成 24 年 4 月～平成 27 年 3 月までの間に、特定の手術を受けた症例おおよそ 150 万人のデータであり、集計期間は平成 24 年 4 月から平成 28 年 3 月までであった。必要とする情報は、エピソード単位で集計された、病名情報とレセプト実績情報であった。抽出にはおおよそ 72 時間が必要であり、抽出されたテーブルのサイズは総計 33GB であった。一時解析の結果取得に必要となった期間はおよそ 1か月程度であった。データを使用した研究者は過去に SQL・R を使った経験があり、今回は参考資料を利用しながらエピソード単位テーブル、レセプト単位テーブルから、データの抽出・集計を SQL で行い、その結果をもとに R を使って解析を行った。

5. 考察

今回作成したシステムで、150 万人～500 万人程度の規模の症例情報をもとに、研究者が必要とする任意条件での症例に関し、情報粒度で情報をサマリーすることが可能であることが示された。また、オンサイトセンターおいて研究者に与えられた環境では、それらの情報を取得するには数日のバッチ処理が必要であることが明らかとなった。バッチ処理完了後のデータを医療系の研究者が解析すると、SQL などの使用経験がある研究者では 1 月程度、使用経験のない研究者でも 2 月程度で一次解析結果までを取得できる可能性が示唆された。

厚生労働省はオンサイトセンターの外部利用を行うための規約整備などを進めており、一つの研究テーマに対し約6か月のオンサイトセンター利用期間を設けることを検討している[3]。今回作成したアプリケーションを用いれば、上記期間内でも東大オンサイトセンターである程度の解析結果を得ることができると考えられる。また、今回作成したアプリケーションを運用することで NDB を使用するうえで、研究者に求められていた技術要件を緩和できると期待される。

今回のケースレポートでの限界として、まず、使用実績が少なく一般化可能性が低い点があげられる。そのため今後は、ユースケースを増やしながら、標準的な利用パターンや限界などを明らかにする必要がある。当該アプリケーションは当面の間、東京大学 NDB オンサイトセンターリサーチユースコンソーシアムの教員および大学院生が、バグ発見・修正などの協力を対価に共同利用する予定である。今後さらに利用ケースを増やすことを検討している。また、データハンドリング過程においてバグが含まれないかの検証も、研究者とともに行っていく必要がある。当面の間、東京大学 NDB オンサイトセンターリサーチユースコンソーシアムの参加者は当該アプリケーションが整形データの質を完全に保証していないことを理解した上で個別研究に用いていく予定である。

今回の事例からも推察されるように、SQL を過去に利用したことのないユーザーでは、一次解析結果を取得するまでに、時間を要した。今後、それらのユーザーを対象とする教育プログラムの構築なども求められる。

6. 結論

オンサイトセンターにおいて NDB を利用するにあたり、研究者側のデータハンドリングコストを軽減するためのアプリケーションを作成した。当該アプリケーションは研究者のユーザビリティを高める可能性があるものの、今後さらなる使用実績に積み重ねと検証が必要である。

7. 謝辞

本研究に関連し、多大な協力をいただいた東京大学 NDB オンサイトセンターリサーチユースコンソーシアムの皆様に深く感謝申し上げます。

参考文献

1) 松居 宏樹，佐藤大介，大江 和彦．レセプト情報等オン

サイトリサーチセンターにおける NDB データの利用~システム環境とNDBの特性に関する報告~,第35回医療情報学連合大会シンポジウム,2016.11.23,神奈川県,パシフィコ横浜
2) 松居 宏樹, 大江 和彦.レセプト情報等オンサイトリサーチセンターにおける NDB データの利用から~操作性,活用可能性,その限界について~,第 35 回医療情報学連合大会シンポジウム,2015.11.2,沖縄県,沖縄コンベンションセンター
3) 厚生労働省. オンサイトリサーチセンターにおけるレセプト情報・特定健診等情報の利用に関するガイドライン(厚労省)(案), 第 38 回レセプト情報等の提供に関する有識者会議, 2017.9.10

表1:オンサイトセンター利用時に行う解析とそれに必要な技術要件

解析の中身	統計/機械学習の知識	プログラミングスキル	ソフト利用(R, SAS, etc)	DB(SQL)	レセプトに関する理解
大規模個票データの解析	○	○	○	○	○
大規模個票データからのサンプリングデータ・集計データの解析			○	○	○
抽出済みデータの解析			○		○
集計データのみを利用					○