厚生労働科学研究費補助金（政策科学総合研究事業（政策科学推進研究事業））
総括研究報告書

レセプト情報・特定健診等情報データベースの利活用の推進に関する研究

研究代表者 大江和彦 東京大学医学部附属病院企画情報運営部　教授

研究要旨

レセプト情報・特定健診等情報データベース（NDB）は、平成 21 年から収集され、現在 90 億件のレセプトが格納されている。しかし、大規模データの処理、学術研究に必要な精度管理、個人情報の取扱等課題は多い。利用には分野横断的な専門性が求められ、大規模データベースであるがゆえに、データのハンドリング自体が研究者にとって極めて難しい上に、そこから得られる知見の可能性を一般研究者が認識できておらず、潜在的な研究ニーズを発掘し、新たな研究着想、利活用着想を支援するためにも NDB 可視化環境の提供も必要である。本研究では、これらの課題を共有し改善方法を検討するため、平成 27 年度は、①NDB の特別抽出データの利活用環境に関する検討、②NDB 基本データセットの利活用に関わる課題調査、③諸外国（米国、韓国）のレセプトデータ（Claim Database）のデータ提供と利用環境の調査検討、等を実施する。

NDB のデータの規模の大きさから生じる「ビッグデータを研究室レベルで扱う困難さ」に研究者は直面しつつある。これを改善するには、柔軟で効率的な大規模計算機資源の活用体制、基本データセットでさえも抽出条件等で柔軟で制約緩和が必要であることが示唆された。これらの解決方策として、韓国で始められた学会と共同で検証した患者サンプルデータセットの考え方、またデータを直接入手しないで計算機資源をネットワークで利用しない米国 VRDC のあり方は参考になると考えられる。

研究分担者氏名・所属機関名　職名

今中雄一・京都大学大学院医学研究科
医療経済学分野　教授

満武巨裕・一般財団法人医療経済研究・
社会保険福祉協会　医療経済
研究機構副部長

研究協力者:

佐藤大介・東京大学医学部附属病院企画情報運営部 助教

大坪徹也・京都大学大学院医学研究科医療経済学分野 助教

國澤進・京都大学大学院医学研究科医療経済学分野 講師

A.研究目的

レセプト情報・特定健診等情報データベース（NDB）は、平成 21 年から収集され、現在 90 億件のレセプトが格納されている。1 カ国の医療機関の 99.9%から収集される悉皆データベースは世界

で類がない。H23 年から試行的、H25 年から本格的に第三者へ提供が開始された（現在まで 40 件）。NDB の利活用に関する研究は、海外のデータセット、オンサイトセンタ（OSC）運用形態、個人ＩＤ精度の限界を明らかにし、OSC の設置、個人 ID 精度に関する情報提供に活用されてきた。レセプト情報等を安全に利用できる OSC が東大と京大に整備され、利用者の増加が見込まれている。

しかし、大規模データの処理、学術研究に必要な精度管理、個人情報の取扱等課題は多い。利用には分野横断的な専門性が求められ、大規模データベースであるがゆえに、データのハンドリング自体が研究者にとって極めて難しい上に、そこから得られる知見の可能性を一般研究者が認識できておらず、潜在的な研究ニーズを発掘し、新たな研究着想、利活用着想を支援するためにも NDB 可視化環境の提供も必要である。

わが国独自の NDB の利活用推進のための分野横断型の研究は十分には議論されておらず、データ解析環境、研究手法、システム処理工程、本データ精度、一般研究者の潜在的ニーズ、などの多くは不明なままである。

そこで本研究において初年度の H27 年度は、①NDB の特別抽出データの利活用環境に関する検討、②NDB 基本データセットの利活用に関わる課題調査、③諸外国（米国、韓国）のレセプトデータ（Claim Database）利用環境の調査、等を実施する。

B.研究方法

１）NDB の特別抽出データの、データベースとしての有効活用に関する問題点と改善点の検討：　研究分担者が NDB より特別抽出として、2016 年 3 月まで NDB データの提供を受けた。このデータ解析実施期間中に生じた問題点のうち、データベースを効率的に利用するにあたっての問題点を記録し、その解決策を考察した。
また、同研究分担者の研究室サーバー環境のセキュリティーの評価：　DB 特別抽出に関してセキュリティーを確保した運営を行っているサーバシステムにおける自営（オンプレミス）運用でのセキュリティーの脆弱性について、専門機関に委託し診断を行った。

２）基本データセットの利活用に関する課題を、脳血管疾患を発症した患者の診療プロセスとアウトカムの関連分析をする研究目的で研究代表者が申請手続きを経て受領したプロセスを元に、①抽出項目の設定方法、②抽出プログラム、③データ精度、の観点から検討した。

３）諸外国の Claim Database の利用環境提供状況の調査のため、日本と類似の国民皆保険制度およびレセプト審査・支払い方式を導入し、一昨年から National Patient Sample という患者サンプルデータの試行提供を開始した韓国、および米国 CMS(Center for Medicare and Medicare Services)は、VRDC(Virtual Research Data center:バーチャル研究データセンター) というバーチャルアクセス機能の提供状況について調査し

た。

C.研究結果

１）①受け取りデータ格納、元データからの抽出：特別抽出申出に際して、CSV ファイルを特殊な圧縮プログラムで圧縮された、1,000 個以上にわたるファイルを受け取っている。これらを個別に解凍し、読み込み、RDB（最も解析に利用しやすいと考えられるデータベース形式）に格納するのに、かなりスペックの高いサーバーでも 1 か月以上かかる。全国データになるとさらに多くなる見通しである。このように RDB 格納に要する時間が膨大である点が大きな研究開始時の障害である。
②受け取りデータを加工し、解析用に抽出するためのサーバーとして、全国規模の解析を行う際データが大量となるが、ネットワークに接続しないローカル機器をあらかじめこのために準備するのは、研究者にとって事前想定不能な資源準備が必要であるため研究開始時の障害となる。
③大きな計算機資源（計算能力とストレージ）を研究室単位で必要とし、研究室だけで一時的にその計算機資源を持つことは困難であった。
④セキュリティー面については、Windows サーバーの Windows Update の遅滞に起因する脆弱性、サーバーの BIOS レベルでの管理ポートに起因する脆弱性が、緊急性の高い事項として指摘をされた。いずれも速やかに対応可能であった。重大な脆弱性は見つからなかった。

２）基本データセットの利点として、3 年間のパネルデータとして利用可能、診療行為や医薬品など 256 項目まで指定した抽出が可能、分析容易なデータ形式でデータを受領可能という点が挙げられた。短所として基本データセットの抽出上限が 256 項目のため、抽出項目は制限せざるを得ない点が挙げられた。

基本データセットの抽出にはプログラム処理が別途必要であることが明らかとなった。

データセットの精度・基本統計量については、今回抽出条件を工夫したにもかかわらず、推計患者数は必ずしも妥当ではなかったが原因は多岐にわたり、不明な点も多かった。

３）昨年から韓国の HIRA-NPS は 5 種類のテーブルで構成されるようになった。具体的には、国家患者サンプル（HIRA-NPS）に加えて、国家入院サンプル（HIRA-NIS）、国家高齢者（65 歳以上）サンプル（HIRA-APS）、および小児患者サンプル（HIRA-PPS）が追加された。追加は、NPS データに確保されていないグループの研究をサポートするために、利用可能とした別々のサンプルデータである。

米国の CMS の VRDC は、研究目的のために CMS のデータにアクセスし、分析するための新しいソリューション（ツール）である。VRDC は研究者がアクセスし、事実上、研究者のワークステーションや PC から CMS データの独自の操作・分析を行うことができる。

D.考察

１）①特別抽出における課題の改善

データ提供（受領）形式を RDB データベース形式とするか、利活用者が指定する圧縮形式とすることにより、受領者がより容易かつ効率的に自身のデータ解析環境にデータ展開できる。

②計算機資源として利活用者がネットワークに接続しないローカルで本利活用専用の計算機資産として保有する資源だけを活用して解析できることを前提とするには、データの規模が大きすぎる。一定の条件を満たすクラウド計算機資源、大学内の高速計算機資源などを活用できるようにすることで劇的に改善すると考えられる。実際、ゲノム解析センターでは高速計算機資源を共用することが当然になっている。

２）基本データセットの長所をさらに生かすためには、抽出条件項目の数を大幅に増やすことと、抽出後のデータ確認やサブセット作成のためのプログラムライブラリを整備することが必要であろう。またデータの精度や学術的利活用の観点からも基本データセットの制約条件について見直しを検討する必要性が示唆された。

３）韓国の HIRA-NPS は 5 種類のテーブル、および米国の CMS の VRDC は今後の NDB の提供と利活用体制のありかたに示唆を与える。

E.結論

NDB のデータの規模の大きさから生じる「ビッグデータを研究室レベルで扱う困難さ」に研究者は直面しつつある。これを改善するには、柔軟で効率的な大規模計算機資源の活用体制、基本データセットでさえも抽出条件等で柔軟で制約緩和が必要であることが示唆された。これらの解決方策として、韓国で始められた学会と共同で検証した患者サンプルデータセットの考え方、またデータは直接入手しないで計算機資源をネットワークで利用しない米国 VRDC のあり方は参考になると考えられる。

F.健康危険情報

該当なし

G.研究発表

1) 「基本データセットの提供について」、第 29 回レセプト情報等の提供に関する有識者会議（平成 28 年 3 月 16 日）、http://www.mhlw.go.jp/file/05-Shingikai-12401000-Hokenkyoku-Soumuka/0000117367.pdf
2) 満武巨裕：レセプトビッグデータ解析の現状と将来. 実験医学,34(5)：799-804,2016.
3) 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,2016.11.2,沖縄県宜野湾市.
4) 大江和彦：わが国の保健医療データベース利活用の現状と今後.第 51 回日本循環器予防学会学術集会,大阪大学中之島センター佐治敬三メモリ

アルホール,2015.06.26,大阪市.
5) 大江和彦:医療における ICT の現状と展望.第 29 回日本医学会総会 2015 関西「医療と IT-近未来の医療はこう変わる-」, 2015.04.11,京都.

H.知的所有権の取得状況

該当なし