厚生労働行政推進調査事業費補助金
（政策科学総合研究事業（政策科学推進研究事業））
総合研究報告書

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 100 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。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、等を実施した。また平成 28 年度は同年度から提供が開始された NDB オープンデータとともに、その情報提供機能の充実と利用に関して潜在的に存在していると考えられる課題も含めてとりあげ、その解決方法についても検討した。具体的には、①ID による同一患者の名寄せ手法、②信頼できる傷病名情報の取得方法、③NDB 特別抽出データの取得と処理方法の効率化、④基本データセットの作成方法、⑤NDB オープンデータの役割および利用例としての国際比較データの作成、の 5 点について検討した。

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、柔軟で効率的なクラウド環境を含めた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。また、データ提供にかかる時間の短縮、受け取ったデータの再 DB 化が効率的に実現できるように、研究者が利用するデータベースエンジンを想定したデータ提供も必要で、現状のテキストデータでの提供には限度がある。また傷病名の精度確保に関する研究をさらに進める必要がある。一方、別の解決方法として、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た。

NDB では、レセプトデータが患者ごとに連結されず、医療機関 x 保険者 x 個人 x 受診年月、に分断されたデータ単位として匿名化されているためにこれらの分断データ単位を再結合することが非常に困難になっている。匿名化 ID が保険者間での結合できるように、全レセプトデータを管理する機関が正確な被保険者台

帳を管理運用することであろう。さらに今後も、NDB からのデータ提供はとしての特別抽出、サンプリングデータセット、基本データセット、NDB オープンデータをはじめとする情報提供機能の充実と利用が必要である。

以上のように課題は残っているが、患者 ID の取扱い、オープンデータの利用、特別抽出データの利用を含む NDB データ活用につき、具体的課題と今後の対応のあり方を示した。今後の NDB データの提供側、およびデータの利用側にとっても、活用推進に役立つと思われる。また NDB オープンデータから生成できるデータを使用した国際統計報告は、諸外国の行政関係者や研究者への波及効果も生み出し得るものであると考えられた。

＜研究分担者＞

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＞

國澤　進　(京都大学大学院医学研究科医療経済学分野　講師)

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)

佐藤　大介　(東京大学医学部附属病院企画情報運営部　助教)

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 100 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。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。

そこで本研究では、これらの課題に関す

る情報を共有し、その改善方法を検討する。

B 研究方法

本研究において 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、等を実施した。

また H28 年度は、①ID による同一患者の名寄せ手法、②信頼できる傷病名情報の取得方法、③NDB 特別抽出データの取得と処理方法の効率化、④基本データセットの作成方法と課題、⑤NDB オープンデータの役割および利用例としての国際比較データの作成、の５点について検討した。

C. 研究結果

H27 年度と 28 年度の結果を関連事項をまとめて記載する。

１）ID による同一患者の名寄せ手法に関する検討：

「全患者 ID1 数」に占める「患者 ID2 を一つもつ患者 ID1 の数」は、63%であり、「全患者 ID2 数」に占める「患者 ID1 を一つもつ患者 ID2 の数」は、68%であった。これにより、氏名の入力ゆれや改名の発生頻度は、保険者番号・記号番号の入力ゆれや保険の変更の発生頻度よりも低いものの、小さいものではないといえる。開発した統合 ID 生成アルゴリズムを適用した結果、患者 ID1 と患者 ID2 の組合せ数に対する統合 ID 数は 42%となった。

２）信頼できる傷病名情報の取得方法の検討

2012 年のある４ヶ月間での入院歴のないレセプトをランダムに４０名分調査したところ、レセプト１件あたりの傷病名数は平均１５．２個であり、レセプトに登録された診療行為と傷病名の適用関係データベースを利用した診療行為から傷病名にウエイトをつけて評価した先行研究では、ほぼ疑いなく患者に存在する傷病名の推定は９３％の正確さであったとの報告があり、連続する数ヶ月のレセプトを処理することで性能が上がる可能性が指摘されている。機械学習と上記手法とを組み合わせた傷病名推定を行うことにより、精度の高い傷病名推定ができる可能性があり、入院患者レセプトについて同一患者で連結した最長６ヶ月のレセプトの傷病名、診療行為、診療科、時間軸で出現情報などを入力とし、正解データとしてＤＰＣの最も医療資源を消費した病名を用いて機械学習することが考えられた。

３）NDB の特別抽出データの利活用環境に関する検討

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。
⑤データ受け取り申請から処理については、ボトルネックになっている行程が、大きく 3 つあると考えられた。 ｉ）有識者会議の承諾から承諾通知発出の間 ⅱ）データ返却、或いは承諾通知発出からデータ抽出作業開始の間 ⅲ）データ抽出作業完了から提供データ格納用外付けハードディスク到着の間。
これらに対し、ｉ）は厚生労働省の事務処理、ⅱ）は厚生労働省およびデータ抽出作業担当も含めた全体的なスケジュール調整・管理、ⅲ）はデータ抽出作業担当からの連絡を改善することによって、それぞれ解消が図られると考えられた。

４）基本データセット：

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。そこで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。

たとえば利用者は、診療行為、医薬品コードの指定にあたり、診療行為マスタは約 6,700 種類、医薬品マスタは約 20,000 種類の中から設定しなければならなかった。このように、利用者が分析上条件設定として必要とする診療行為や医薬品は 256 種程度よりはるかに多いことがわかる。これを解決するには、各診療行為や医薬品をカテゴリーに抽象化（指定粒度を意味的に粗くする）する必要がある。医薬品の場合、一般名化、WHO-ATC 分類などの粒度に抽象化する、などの手法が必要である。

一方、基本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。

５）諸外国のレセプトデータ利用環境の調査：

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状況であった。

６）NDBオープンデータ：
NDBオープンデータの役割、および利用例としての国際比較データの作成に関する検討
都道府県ごとに、高齢者人口を医療需要総量を反映するものとみなして、どれだけの提供量があるかを示した。これらの指標は、治療へのアクセスの指標ともなりうると考えられる。一方で、都道府県内でも、地域間の格差が大きく、これらの指標は、各県内の格差の大きい地域を足し合わせた平均値である点に、特に留意すべきである。指標の解釈の例もコメントとして付記しているが、解釈も一例であることに留意が必要である。

７）NDBオープンデータの利用による国際比較データを作成
CTの対1000人当たりの施行件数、MRIの対1000人当たりの施行件数について、国際比較を事例としてOECD（経済協力開発機構）が公表している諸外国の医療の質データに関連した日本の新しい指標作成について検討した。OECD加盟国間で比較可能なデータを作成できた。

D. 考察

１）IDの連結課題：レセプトデータが患者ごとに連結されず、医療機関x保険者x個人x受診年月、に分断されたデータ単位として匿名化されているためにこれらの分断データ単位を再結合することが非常に困難になっている。この問題を解決するには保険者が厚労省にデータを匿名化して出す場合に、医療機関x個人x受診年月についてはその同一個人ごとに１匿名化 IDとなるように処理するとともに、その匿名化IDが保険者間での結合できるように、全レセプトデータを管理する機関が正確な被保険者台帳を管理運用することであろう。

さらに今後も、NDBからのデータ提供はとしての特別抽出、サンプリングデータセット、基本データセット、NDBオープンデータをはじめとする情報提供機能の充実と利用が必要である。

２）傷病名の精度の課題については、機械学習により傷病名推定を行うには、大量の正解データが必要で、入院レセプトではDPCの主たる医療資源を消費した病名以外には、退院時サマリの病名を電子的に収集する手法が考えられる。外来レセプトにおいて正解データである傷病名をどのように入手するかについてはさらに検討が必要であるが、特定の傷病名で実施される診療行為と、そうでない診療行為のリストを作成して分析することが必要であろう。

３）特別抽出における課題の改善では、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。研究者が利用するデータベースエンジンを想定したデータ提供も必要で、現状のテキストデータでの提供には限度がある。また傷病名の精度確保に関する研究をさらにすすめる必要もある。

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ことも参考にすべきである。

４）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。各診療行為や医薬品をカテゴリーに抽象化（指定粒度を意味的に粗くする）する必要がある。医薬品の場合、一般名化、WHO-ATC 分類などの粒度に抽象化する、などの手法の採用も必要である。

５）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。

６）オープンデータセットの利用については、都道府県内には地域間の格差が大きくモザイク状態になっているため、得られる指標は、各県内の格差の大きい地域を足し合わせた平均値である点に、特に留意が必要である。このように解析には限界を含めた解釈が併記されることが望ましい。H28 年度分担報告ではその解釈の例としてコメントも作成したが、解釈もあくまで試行的なものであることも留意が必要である。

７）NDB オープンデータを利用した国際比較指標の作成

例えば、2015 年に OECD が提出を求めている 129 項目の HQ において、日本は 53 項目を提出しているが、加盟 35 ヵ国中 32 位と低い。ちなみに OECD 加盟国の平均提出件数は、89 項目である。そのため、日本はデータ提出状況を改善することが望ましい。NDB オープンデータから今回新しいエビデンスを作成することができたが、日本の全人口の保険医療記録から作成されたデータは、国内だけではなく諸外国に向けた情報としても価値のあるものと考えられる。

E. 結論

考察で述べたように課題は残っているが、患者 ID の取扱い、オープンデータの利用、特別抽出データの利用を含む NDB データ活用などの課題につき、具体的課題と今後の対応のあり方を示した。今後の NDB データの提供側、およびデータの利用側にとっても、活用推進に役立つと思われる。また NDB オープンデータから生成できるデータを使用した国際統計報告は、諸外国の行政関係者や研究者への波及効果も生み出し得るものであると考えられた。

F. 健康危険情報

特になし。

G. 研究発表

1. 論文発表

1) 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.
2) 満武巨裕、大江和彦、今中雄一:NDBオープンデータを研究利用に活用する:医療技術(CT,MRI,PET)の利用に関する国際比較の試み、社会保険旬報,第2661巻:12-16,2016年
3) 大江和彦. 医療情報データベースの基盤整備, 情報管理. 2016, vol. 59, no. 5, p. 277-283.

2. 学会発表

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) 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.
3) 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.
4) 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.
5) 満武巨裕(「諸外国の医療ビッグデータ」、第2回データヘルス時代の質の高い医療の実現に向けた有識者検討会(平成28年5月23日)
6) 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、今後の第三者提供の方向性について 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,医療情報学連合大会論文集36回1号,138-140(2016.11).
7) 佐藤 大介, 大江 和彦,医療データベース利活用の国内基盤の最新状況,レセプト情報等オンサイトリサーチセンターの試行的利用について,日本薬剤疫学会学術総会抄録集,巻22,p51 (2016.11).
8) 大江 和彦,医療ビッグデータとこれからの医療,日本臨床検査自動化学会会誌(0286-1607)41巻4号,371(2016.08).

H. 知的財産権の出願・登録状況

該当なし