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））

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

平成 27 年度　総括・分担研究報告書

研究代表者　大 江 和 彦

平成 28 (2016) 年　3 月

目　　次

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

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））
分担研究報告書

レセプト NDB（ナショナルデータベース）特別抽出データ利活用推進のための課題

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)
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)

要旨
目的： NDB の特別抽出データを有効に利用するための環境を提言していく
方法： NDB の特別抽出データを受け取り、データベースとして解析を行った。この間に生じた問題点と改善点を検討する。また抽出データの受入れの内部環境要件を検討するにあたり、研究室データベース環境のセキュリティー診断を実施した
結果・考察：　１）NDB の特別抽出データの解析について課題が存在する。
○ 受取りデータ形式が特殊でかつテキストを RDB に格納するまで膨大な時間(1 か月以上)がかかりうる　→　Microsoft SQL の RDB 形式でのデータ渡しにより、利活用の効率が大幅に向上する。
○ 受け取りデータが膨大であり、その解析のためのデータサーバーが膨大に必要、しかしデータ整理後はそれらが不要になる　→　必要なセキュリティを確保したデータ部分のクラウドサーバーの活用により、必要なセキュリティを確保した利活用の効率が大幅に向上する。
○ 内部環境のみで解析まですべて終了させることが不可能、GIS システムやスーパーコンピューターでの解析が必要　→　万全のセキュリティを確保できるように単純なデータに落とし込んだものを解析環境に持ち出せることにより、必要なセキュリティを確保した利利活用の効率が大幅に向上する。
２）当研究室のデータ管理は、情報セキュリティーマネジメントシステム適合性の第三者審査登録機関による認証を取得して維持されている（国際規格 ISO/IEC 27001:2013, 国内規格 JIS Q 27001:2006）。この上で、ソフトウェア管理によるセキュリティーの脆弱性の存在を、専門機関に委託し診断したところ、重大な脆弱性はなく、継続的なセキュリティーの確保を行うことで、物理的にもソフトウェア的にも高いセキュリティーの維持が可能とみなされた。
結論： NDB データを有効に活用するためには、データの受け渡し方法から解析環境まで柔軟な対応が求められる。また、必要なセキュリティー対策を講じていることで、大学内研究室レベルで安全性の高いシステムが維持できる。

A．目的

NDB の特別抽出データを有効に利用するための環境を改善するための条件を検討し、改善のための施策案を提示することを目的とした。

B．対象・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

当研究室では、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。このデータを用いて解析を行い、報告を行っている。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。

2）研究室サーバーという環境のセキュリティーの評価

当研究室では、情報セキュリティー管理方針は、情報セキュリティーマネジメントシステム適合性の第三者審査登録機関による認証を取得して維持されている（国際規格 ISO/IEC 27001:2013, 国内規格 JIS Q 27001:2006 認証登録番号 IS75998）。NDB 特別抽出に関しては別途セキュリティーを確保した運営を行っているが、今回、そのほかのサーバーシステムについて、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。

脆弱性診断を行う期間は、複数の候補を挙げ、最終的に株式会社ラックへ発注した。

診断は 2016 年 2 月 22 日、23 日にかけて行い、大学のファイアーウォールのさらに内部での脆弱性を診断するため、研究室内で直接ネットワークに接続し診断を行った。

実施された診断には下記が含まれた。

1．ポートスキャン：稼働サービス特定を実施し

2．市販脆弱性ツールによる診断：

(サービス、ミドルウェア、OS 等に存在する脆弱性を調査)

3．独自ツールによる診断および 2 の結果確認：弊社独自ツールを使用し②の診断では検出できない HTTP や SMTP 等の問題を診断、および市販ツールで検出された問題が存在するか確認

C．結果

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

I

受け取りデータ格納、元データからの抽出

現在特別抽出申出に際して、CSV ファイルを EXE（特殊なプログラム）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り、研究実施スケジュールの遂行に多大な支障をきたす。

そこで、データ利用期間内に確実に上げるため、以下の二つを提案する

改善案１　　RDB のデータベースファイルとして提供を受ける

具体的には、研究室の指定する RDB のデータベースファイル形式で提供を受ける。例えば SQL Server 2014 Enterprise Edition に付随するクラスター化カラムストアインデックスによるデータベースの圧縮を施したデータベースを、データベースファイルとして提供受けることができれば、研究室でのデータ運用は、アタッチするだけで開始でき、データベース容量は非常に小さく受け渡しも簡便になり、その検索能力も高くなる。

※　データ提供としては、SQLDB と CSV 両方をもらうのが最良

改善案２　　受領データの格納形式や圧縮方法について、申請者が指定できるようにする。

現在の仕様､圧縮 EXE ファイル以外での提供が無理な場合､圧縮 EXE ファイルの一覧表と、それぞれに圧縮されているファイル一覧（内容を含む）を提供いただく

II　セキュリティーを確保しながらより合理的な解析環境を構築するための提案

特別抽出データは、申出者が管理権限をもつ物理ディスクにデータを格納し、外部ネットワークとは断絶した環境内で運用することでセキュリティーを確保するとされている。

一方、受領データの運用に必要なハードウェアのスペックはデータ受領前では未知であるため、複数年度で全国にわたる大規模データを研究に要する場合、事前に必要な器材を選定・調達することは極めて困難となる。特に、受領データ後にスペック不足が明らかとなった場合、追加の器材を調達するための予算は直ちに確保することは極めて困難となる。

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なり、研究室で用意できる環境では、能力の不足が起こり得る。従来、独立したネットワークにつながらないサーバーのみでの運用をセキュリティーの高い状態として考えられているが、この状態を維持するだけでは能力を増強するのが物理的にも予算的にも困難となってくる。

このことを改善するため、以下を提案する。

改善案３　クラウドコンピューティングの活用

解析用のデータベースを構築するまでに、大量のデータを展開する場所と、そのための高スペックなサーバーが必要となる。いったんデータベースを構築した後にも別途高性能のサーバーが必要となるが、必要なスペックが異なり、用途ごとに高性能なサーバーを購入し、設置・運営するのが非常に困難になる。クラウドコンピューティングでは、データ整理や、多変量解析など、その用途に応じてサーバースペックを可変できるため、柔軟で効率的な運用が可能になる。さらに、クラウド内でのみの解析運用を行う場合、データを一切「外」へ出す必要がなくなり、かえってセキュリティーの高い運用が可能になるとも考えられる。ただし、この信頼性はクラウドコンピューティングを提供する会社の信頼性に依存する。

改善案４　通常運用のサーバーによる取り込み、処理を許可してもらう、処理後は、データベースを独立した運用に移行する

従来の申請では、データ自体を特別に指定した独立した機体に格納することとしている。この方法では、目的とする能力が場面場面で異なり、いずれにおいても十分な性能で作業ができない。つまり、データの格納に必要なサーバー

（主に作業スペースなど）と、解析に必要なサーバー（主に計算能力）など、解析全体を通じて必要なサーバーが異なっている。

そこで、特にデータ整理を行う期間、通常ほかで運用している研究室内のサーバーを一時的に NDB 用に割り当て、研究用データベースを確立し、その後、独立した状態に戻す、という柔軟な運用を提案する。

具体的には、データ自体を NDB 専用のサーバー（NDB ストレージと呼びます）に置き、データアクセス、処理の命令を行うサーバーと分けます。この「頭」となるサーバーを、通常運用のサーバーから一時的に流用し、NDB ストレージ内のデータを整理し、整理後、頭を切り離す、という流れができる。

この利点は、有限な資源を効率的に利用できるほかに、「頭」は NDB ストレージを切り離した状態では通常運用ができるため、セキュリティパッチのアップデートなど、重要なメンテナンスがスムーズに行えることもある。

つまり、物理的なつながりを遮断する旧来の方法にこだわることなく、システムとしての接続可能性に依った運用を柔軟に行うことで、資源を有効利用できることが利点になる。

この前提としては、セキュリティーの高いサーバーの運用が確保されている必要がある。

改善案５　サーバーの概念の見直し

現在の運用では、物理的な違いを持って、サーバーが違うのでアクセス権が違うことを明確にしてる。具体的には、甲サーバー（元データ）と乙サーバー（解析データ）を物理的に分けて管理するなどを示している。しかし、いずれも高性能の処理が求められ、本来であれば能力を補完して運用するものになる。これは前述の改訂案２と同様の内容である。

このことを解決するために、サーバーに対する管理というものを、機能単位で明示し、そのアクセス権を制御することでセキュリティーを保つようにできる。

具体的には、1 つの物理的サーバー内（DB1）に、甲サーバーと、管理の異なる乙サーバーを構築し、それぞれ適切なアクセス権を付与します。例えば Microsoft では、Domain Controller と呼ばれる機能により、アクセス権の集中管理を行い、解析端末でのアクセス権との整合性を保つことができる。

このシステムでは、従来甲・乙、独立した 2 台のサーバーを、上記 DB1 と同等の機能を持つ DB2 を並列に処理させることで、どの処理についても約 2 倍の処理能力を得ることが可能になる。

III

解析環境の問題点

独立系の PC での解析の限界

最近では、解析対象のデータ量が膨大になってきている。また、解析手法自体も洗練されてきており、それに伴い必要な処理が高度化してきている。

データ量：単純に地域やデータ内容が増えればその分増加する

解析方法：単純な回帰モデル　→　マルチレベル

高機能 PC でも 1 昼夜必要

また、ネットワーク分析、GIS 分析など、専用の解析ソフトやサーバーが必要になってきている。

このため、独立系の解析環境に、大量データを扱える機材と、そのソフトウェアを準備することが、現実の予算的に困難となる。例えば GIS 解析を行うために、その解析サーバー自体が必要になり、一つ一つのソフトウェアが非常に高額となる。

また、実際には、マルチレベル分析は、スーパーコンピューター（大型計算機）による分析

が必要になるなど、いわゆる外部（通常運用場所）での解析が必須となる

このため、現行では独立したネットワーク、空間での解析のみではなく、次のように改定を行うことで、より迅速で有用な解析を実施できると考えられる。

改善案６　データを個票レベルであっても、単独では意味のないデータに落とし込み、通常環境での解析を行う

具体的には、データの加工はすべてサーバー室内（独立した環境）で行い、解析直前のフラグとなったデータのみを、通常環境へ移行し、解析を行う。

イメージとして、図表 1 を参照されたい。このイメージは、実データにまったく関係なく作成したデータであるが、実データで作成しても同様に、単独ではまったく意味を持たないデータとして作成が可能である。このイメージ図では説明をわかりやすくするため変数に意味が若干推測しやすくしているが、変数名を 2 文字以内にするなどの制限により、さらに単独ではわかりにくいデータとして規定することができる。

集計され、許可されたデータのみを持ち出すのではなく、個人情報を全く含まず、かつ単独で意味をなさない加工されたデータを、持ち出し、通常環境で解析を行い、その解析結果の解釈を含めて初めて、意味のあるデータとしての持ち出し許可を判断してもらうことでもセキュリティーを確保しつつ運用が可能と考えられる。

２）研究室サーバーという環境のセキュリティーの評価

研究室で運用されるサーバーおよび PC をサンプリングし、診断を行った。サーバーは、実サーバーのほか、仮想サーバー、およびサーバー管理ポートも診断に含めた。

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を行った。そのほか中等度以下の脆弱の可能性としての指摘を受けたが、必要に応じて随時対応を行っている。重大な脆弱性は見つからなかった。

D．考察

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

NDB データの性質上、どれだけセキュリティーを確保できるかに焦点がおかれ、非常に制約の高い状況での解析が求められている。今後、ますます高度な解析が求められるようになる中、セキュリティーが確保できる中での、柔軟な運用が期待される。

２）研究室サーバーという環境のセキュリティーの評価

継続的なセキュリティーの確保を行うことで、物理的にもソフトウェア的にも、高いセキュリティーの維持が可能であった。

E．結論

大学研究室内の運用にてセキュリティーの高い環境を維持できることが第三者的にも示され

た。
　NDB のデータの効率的な運用のためには、柔軟な対応が期待される。

Ｆ．健康危険情報

特になし

Ｇ．研究発表

特になし

図表１　イメージとして、何らかの解析データにまったく関なく創作したデータ

staindex	sex	agec	outcome	sev	com1	com2	com3	com4	area1	area2
1	2	4	1	1	1	1	1	1	232	232
2	1	4	0	4	0	0	1	0	176	176
3	1	1	1	3	0	1	1	0	52	51
4	1	1	1	3	1	0	1	1	319	321
5	1	2	0	2	1	0	1	1	158	157
6	1	5	0	4	0	1	1	0	165	166
7	2	2	0	4	0	0	0	1	200	201
8	1	5	0	1	0	1	1	1	393	394
9	1	4	1	3	0	0	1	0	419	419
10	2	3	0	1	1	0	0	1	352	353
11	1	2	0	1	1	0	1	1	131	133
12	1	3	0	4	1	1	1	0	180	180
13	1	3	1	4	0	1	1	1	364	363
14	1	5	0	3	1	0	1	0	246	247
15	2	3	1	4	1	0	0	0	267	267
16	2	3	1	1	1	1	0	1	171	171
17	2	1	1	3	0	0	0	1	304	305
18	1	4	0	3	0	1	1	0	130	131
19	2	4	0	3	1	1	1	0	278	279
20	2	1	1	2	1	0	0	0	256	257
21	1	4	0	4	0	0	0	0	382	383
22	2	2	0	2	0	0	1	0	267	266
23	2	4	1	3	1	0	0	1	322	322
24	2	3	0	2	0	1	0	0	88	90
25	2	3	0	1	0	0	1	0	393	394
26	1	2	0	2	1	0	1	0	311	313
27	2	5	1	4	1	1	0	0	361	360
28	2	2	0	2	1	1	0	0	187	189
29	2	3	1	1	1	1	0	1	405	407
30	1	2	1	3	1	1	0	1	221	220
31	2	4	0	2	1	0	0	0	190	189
32	2	1	0	2	1	1	1	1	338	340
33	2	4	0	4	1	1	0	0	156	156
34	1	5	0	2	1	0	1	0	118	120
35	2	5	0	2	1	1	0	0	115	116
36	2	2	1	3	1	0	1	0	293	293
37	1	5	1	3	0	0	1	0	43	44
38	2	5	1	3	1	0	0	1	313	314
39	1	4	1	4	0	1	1	1	216	215
40	2	1	0	2	1	0	0	0	224	225
41	2	1	1	1	1	1	1	0	367	367
42	2	5	0	4	1	0	1	0	420	421
43	2	1	1	3	0	1	0	0	49	51
44	1	1	1	2	0	0	0	1	154	153
45	1	2	1	3	0	0	1	1	127	128
46	2	2	0	3	1	1	0	1	447	448
47	2	4	1	1	0	0	0	0	404	406
48	2	3	0	1	1	1	1	1	64	66
49	1	2	0	4	1	1	1	1	209	210
50	1	3	1	1	0	0	1	1	304	304
51	2	2	1	3	1	1	1	0	301	300
52	2	2	1	1	1	0	1	0	141	141
53	2	1	0	1	0	1	1	1	103	103
54	2	4	1	2	1	1	1	0	166	167
55	2	3	0	2	0	1	0	1	349	350
56	2	1	0	2	1	1	0	0	122	124
57	1	5	0	1	0	1	0	1	113	115
58	2	3	0	4	0	1	0	1	373	373
59	1	4	1	1	1	1	0	1	446	445
60	1	2	0	1	1	0	1	1	131	131
61	2	4	0	1	0	0	1	0	201	203
62	1	1	1	2	1	0	1	0	231	230
63	2	5	0	4	1	1	1	1	358	359
64	1	4	0	3	0	0	0	1	162	162
65	1	1	0	4	0	0	0	0	218	220
66	2	2	1	2	1	0	1	1	450	452
67	1	2	1	1	1	0	1	1	237	239
68	1	4	1	2	1	1	1	1	210	211
69	2	4	0	2	0	0	1	0	295	296
70	1	2	1	2	0	0	1	1	282	282
71	1	5	0	1	0	0	1	0	144	146
72	1	4	1	2	1	1	0	1	263	265
73	2	1	0	4	0	1	0	0	9	9
74	2	2	1	4	0	0	0	0	19	18
75	2	3	0	4	1	1	0	1	173	175
76	1	3	0	4	1	1	0	1	181	182
77	2	2	0	2	1	0	0	1	346	346
78	1	3	1	2	1	1	1	0	116	118
79	2	5	0	2	0	0	0	0	393	393
80	1	1	1	1	0	1	1	1	303	303
81	1	1	0	1	0	0	0	1	268	268
82	2	2	1	4	1	0	1	1	444	446
83	2	2	0	4	0	0	1	0	125	124
84	2	5	0	4	0	0	0	1	66	65
85	1	4	1	3	1	1	1	1	130	132
86	2	2	1	3	1	0	0	1	67	66
87	2	2	0	2	0	0	0	1	289	288
88	2	2	1	2	0	1	0	1	297	298
89	2	5	0	1	0	0	0	0	56	58
90	1	5	0	1	0	1	0	1	321	320
91	1	5	1	4	1	0	0	1	247	248
92	2	3	0	4	1	0	0	0	128	129
93	2	1	1	1	1	1	1	0	437	437
94	2	4	1	1	1	1	1	1	70	70
95	2	1	0	4	1	0	1	0	68	67
96	2	3	0	4	0	0	1	0	275	277
97	2	1	1	4	0	0	1	0	149	150
98	1	5	0	4	0	1	1	0	307	307
99	2	4	1	4	1	1	1	1	39	38
100	1	2	0	2	1	1	0	0	73	75

厚生労働科学研究補助金（政策科学総合研究事業(政策科学推進研究事業）)
分担研究報告書

レセプト情報・特定健診等情報データベースの申出者対応部門の充実

研究分担者　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

研究要旨

本報告書は、今後の日本におけるレセプト情報･特定健診等情報データベース（以下、NDB）の情報提供機能について、諸外国の先進的事例を参考にして、今後の充実について検討する。

今年度は、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を調査対象とした。韓国の患者サンプルデータは、既に台湾において被保険者ファイル(ID)から、100万人をランダムサンプリングし、抽出された被保険者の入院、外来、調剤レセプトデータを提供している例を参考にしている。

韓国は、台湾を参考にランダム抽出した患者の入院、外来、調剤レセプトデータ（HIRA-NPS (HIRA NationalPatient Sample)の試行提供を一昨年から開始した。HIRA-NPSは、韓国国内において1年間に医療機関を利用した全患者対象を母集団として、性別・年齢(5歳単位)区間による患者単位の層化系統抽出を行ったデータセットである。現在、韓国のHIRA-NPSは5種類のテーブルで構成されている。また、患者サンプルデータは、5つの学会との覚書（MOU）を交わして検証が行われた。例えば、主要な検証の一つに、糖尿病およびジペプチジルペプチダーゼ4阻害剤の使用に関する有病率について、患者サンプルデータを用いら推定値が既存研究と整合性があることが証明された。一方、患者サンプルデータの期間が1年間単位であり、個人の縦断突合ができないために、有病期間が長い慢性疾患などの分析にも適していないことが指摘されている。

また米国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を提供して、利用者に効率的かつ対費用効果の高い方法でメディケアとメディケイドプログラムデータへのアクセス環境を提供している。VRDCを利用する研究者は、承認されたデータファイルへ直接アクセスができ、CMSのセキュアな環境の中で研究が実行できる。また、研究者本人のローカルマシン（自身の研究室のワークステーションやPC）に、集約されたレポートと結果をダウンロードすることができる。

日本のレセプト情報等データベース（以下、NDB）から提供されるデータセットは、厚生労働省側で複雑な構造の電子レセプトを研究者の要望に応じて、ある程度分析し易

い形式に加工して提供する特別抽出と一月分のサンプリングデータの提供サービスが存在している。加えて基本データセットの設計と作成が検討され、今年度から試行提供が始まった。日本の NDB は、提供開始（2011 年 11 月）から平成 28 年 3 月までの承諾件数は合計 94 件となったが、台湾と韓国の平成 25 年の提供件数は、韓国は 115 件／年、台湾は 270 件／年である。したがって平成 28 年中に公開される NDB オープンデータをはじめとする集計情報も含めて、量と質の増加が必要である。その方向性の一つとして、韓国 HIRA の患者サンプルデータ、米国 CMS の VRDC は今後の日本の**申出者対応部門を充実**する上で有益な先行事例である。

A. 研究目的

本報告書は、今後の日本におけるレセプト情報･特定健診等情報データベース（以下、NDB）の情報提供機能について、諸外国の先進的事例を参考にして、今後の充実について検討する。

今年度は、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を調査対象とした。韓国の患者サンプルデータは、既に台湾において被保険者ファイル(ID)から、100万人をランダムサンプリングし、抽出された被保険者の入院、外来、調剤レセプトデータを提供している例を参考にしている。

また米国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を提供して、利用者に効率的かつ対費用効果の高い方法でメディケアとメディケイドプログラムデータへのアクセス環境を提供している。VRDCを利用する研究者は、承認されたデータファイルへ直接アクセスができ、CMSのセキュアな環境の中で研究が実行できる。また、研究者本人のローカルマシン（自身の研究室のワークステーションやPC）に、集約されたレポートと結果をダウンロードすることができる。

日本のNDBから研究者に提供されるデータ件数も近年増加傾向にあるが、先行事例の米国、台湾、韓国にはおよばない。したがって平成28年中に公開されるNDBオープンデータをはじめとする集計情報も含めて、量と質の増加が必要である。その方向性の一つとして、韓国HIRAの患者サンプルデータ、米国CMSのVRDCは今後の日本の**申出者対応部門を充実**する上で有益な先行事例である。

B. 研究方法

韓国は、HIRA から患者サンプルデータについての資料提供を基にしている。（内容は、Kim L らの”A guide for the utilization of Health Insurance Review and Assessment Service National Patient Samples” Epidemiol Health.Vol;36,ArticleID:e20140008,2014 に要約されている）

米国は、CMS のデータ提供に関するサポート業務を行っている ResDAC(Research Data Assistance Center)がインターネッ

トで提供している Introduction to the Virtual Research Data Center (VRDC)（URL: https://www.resdac.org/cms-data/request/cms-virtual-research-data-center）を基にしている。

C. 研究結果

韓国は、日本における審査支払機関に該当する HIRA(The Health Insurance Review and Assessment Service（健康保険審査評価院）)が、HIRA-NPS (HIRA NationalPatient Sample)として、韓国国内において 1 年間に医療機関を利用した全患者対象(約 4600 万名)を母集団として、性別・年齢(5 歳単位)区間による患者単位の層化系統抽出を行ったデータセットである。HIRA の患者サンプルデータは、患者の診断、治療、処置、手術歴、および医療サービス研究のための貴重な情報源である処方薬情報を含んでいる。しかし、入院患者の 10%および外来患者の 90%で構成されている国会の国家患者サンプル（NPS）は、重症化した入院患者を調査するのに十分な数を確保していない可能性がある。

そこで昨年から韓国の 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。

しかし、これらの患者サンプルデータは、一年間分のレセプト請求データをソースとして作成されている断面調査である。患者らのプライバシーを保護するために、毎年サンプルデータは作成され、特定の個人または医療サービス提供者も患者サンプルであって横断的な調査はできない。つまり、複数年の患者サンプルデータを使っても患者の長期間の観察研究を行うことができないようになっている。だが、また、医療援助プログラム、政府支出、および退役軍人患者のデータも請求データに含まれている。

しかし、レセプトの複雑な構造とレセプト請求データの膨大な量は、研究者に一定以上の負担を課すことになる。また、膨大なデータ量は、研究を行う上で非効率性をもたらす可能性がある。これらの制限事項を解決し、レセプトデータの利用向上と研究者へのアクセシビリティを向上させるために、HIRA は 5 つの異なる機関によって行われた検証を経た患者サンプルデータを開発した。

患者サンプルデータは、それぞれ 5 つのテーブルで構成されている。すべてのテーブルは、キーID を使用してリンク可能となっている。基本的属性テーブルは、このような性別･年齢および医療援助プログラムといった社会人口学的特性、主要診断名、二次診断名、診療開始日や実日数、患者の自己負担額などのから構成されている。医療サービステーブルは、入院患者のための処置、治療、薬剤情報など、患者に提供され入院と外来医療サービス情報から構成されている。診断情報テーブルは、

患者の診断情報がすべて含まれている。このテーブルは、患者の合併症または全ての病名の履歴が必要と判断された場合に使用される。外来処方テーブルは、成分、投与量と供給日といった、外来患者のための処方薬剤の情報から構成されている。プロバイダテーブルは、患者の受診した医療機関の種類（プライマリケア、二次ケア、専門治療）、位置、病床規模、運営（経営）母体のタイプといった情報から構成されている。

患者サンプルは、韓国の患者全体の代表性を有しており、5 つの学会として韓国予防医学会(Korean Society for Preventive Medicine)、韓国医療経済学会(Korean Association of Health Economics and Policy)、韓国医療情報・医療統計学会(Korean Society of Health Information and Health Statistics)、韓国医療政策・管理学会(Korean Academy of Health Policy and Managemen)、韓国疫学学会(Korean Society of Epidemiology)との覚書（MOU）を交わして、検証されている。したがって、研究を行う際に利用したデータが母集団の特性を有しているかについて検討するための説明を省くことができることが証明されている。主要な検証結果として、糖尿病およびジペプチジルペプチダーゼ 4 阻害剤の使用の評価の韓国有病率について、推定値は人口全体と整合のある患者サンプルであることが証明されている。また、血糖降下薬利用の処方の推定値についても検証が成功した。さらに、それぞれの血糖降下剤の外来処方率はすべて 95%信頼区間内であった。「視力低下や失明に関連する疾患の社会的コスト」、「患者サンプルおよび人口の試験」においても、主要な眼疾患（白内障、緑内障、黄斑変性症、糖尿病性網膜変化）については女性患者においてより高い医療サービスの利用を示した。

一方、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。これまで CMS は、外部メディアに保存した暗号化データファイルを研究者に提供してきたが、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。VRDC は、より効率的で費用対効果の高い方法でタイムリーなデータにアクセスするための安全なメカニズムを提供している。

ただし、ユーザー要件としては、SAS プログラミング言語、ブロードバンドインターネット接続、Java6 以上のローカルマシンへのインストール、MS の Internet Explorer または Mozilla Firefox 、Windows XP またはそれ以降の Windows オペレーティングシステムでなければならない。

複数の研究者が CMS VRDC 内の単一のプロジェクトに取り組むことも可能になっており、彼らの SAS ライブラリ内の仮想デスクトップ内で共同作業することができる。しかし、CMS VRDC のオンライン・セキュリティトレーニングを受け、完了した証拠を提供する必要がある。研究者全員が全てのセキュリティ要件を満たした後、利用者のアクセス権が付与され、CMS VRDC 環境に接続するために必要

なソフトウェアを備えたパッケージが提供される。ただし、複数の研究者が同じプロジェクトで作業している場合 VRDC の利用にあたっては、それぞれの申請および権利を得なければならない。人数分だけシートと呼ばれる利用権利を購入する必要がある。また、シートを共有することもできない。

このような条件の基、利用者は VRDC により次のファイルにアクセスすることができる。

·マスター受益者ファイル

·メディケア・パート A 、 B、D のレセプトデータ

·メディケアプロバイダー分析ファイルと MedPAR ファイル

·メディケイド（ MAX ）ファイル等である。

利用者は、SAS による分析が終了した後、ダウンロードできるのは集計情報や統計情報に限られている。個人を特定できるような情報または保護された健康情報は、VRDC から取り出すことはできない。CMS VRDC からデータをダウンロードするためのすべての要求は、個人を特定できる情報または保護された健康情報をスクリーニングするために、出力審査を通過する必要がある。

ダウンロードファイルは、研究目的にもよるが、地理的な単位（州、郡等）、診断グループレベルに集計しなければならない。出力形式は、Excel テーブル、集約された SAS データセット、SAS 出力ファイル、ワード文書、PDF 文書である。CMS のダウンロードデータに関する審査プロセスは、一般的には 2 営業日（48 時間）以内にとしている。しかし、審査内容が複雑である場合には追加の時間が必要とされる。

D. 考察

NDB データを利用する際、厚生労働省や関係省庁・自治体に属さない研究者等への第三者提供については、有識者会議（レセプト情報などの提供に関する有識者会議）において医療サービスの質の向上を目的とする公益性の高い研究であることが前提で、有識者会議の承諾を得なければならない。承諾の敷居は高く、研究者等への第三者提供を検討した第一回は、43 件の申出に対して承諾件数は 6 件であった。提供開始（平成 23 年 11 月）から平成 28 年 3 月までの承諾件数は 94 件となった。参考までに、日本と同様の社会保険方式でありレセプトも存在する台湾と韓国の 2013 年の提供件数は、韓国は 115 件／年、台湾は 270 件／年である。

この承諾件数が低い原因として、次の点を有識者会議は指摘している。(1)申出者が求めるデータ項目が実際に格納されているデータでは実現困難であった申請が存在した、(2)データ提供にあたっての各種要件や必要な事項を申出者が十分に把握していない申請が存在した、(3)提供側の情報提供が不十分であった等である。

しかし著者は、上記以外に NDB データの利用規約に原因があると考える。第一に、利用者の申請した範囲に分析方法が限定されてしまうことである。つまり、探索的にあれこれと自由に研究することができず、限定されたデータ項目及び期間しか提供されない。また、成果の公表前に、厚生

労働省の承認が必要であり、承認を得なければ発表することができない。加えて、データベースへの複写回数は原則一回、利用場所の施錠と入退室状況の管理、データの持ち出しは原則不可などの規約を守らなければならず、利用場所への外部検査官の立ち入り検査にも応じなければならない。実際に承諾を得た大学や研究所では、大半の研究機関では利用規定を満たすために新たな物理的場所の確保や入退室記録装置を導入しているケースが多く、予算等の問題もあって申請を見合わせる研究者が多い。

ここまで厳格な管理が求められるのも、NDB は医療機関から提供された医療関連情報だからであり、現時点では個人情報に準ずる取り扱いをするということになっているからである。ただし、レセプトに記載されている氏名や住所等の個人情報は全てハッシュ関数による暗号化が施されており、個人を特定することはまず不可能と言える。

NDB のオンサイトセンターは、東京大学と京都大学に拠点がおかれて開始される予定である。しかし、全研究者が二つの拠点に集まらなければならないのは、物理的な距離の問題、分析作業が終了するまでの滞在費用なども問題もある。

NDB から提供されるデータセットは、厚生労働省側で複雑な構造の電子レセプトを研究者の要望に応じて、ある程度分析し易い形式に加工して提供する特別抽出と一月分のサンプリングデータの提供サービスが存在している。加えて基本データセットの設計と作成が検討され、今年度から試行提供が始まっている。平成 28 年中に公開される NDB オープンデータをはじめとする集計情報も含めて、量と質の増加が必要である。その方向性の一つとして、韓国 HIRA の患者サンプルデータ、米国 CMS の VRDC は今後の日本の**申出者対応部門を充実**する上で有益な先行事例である。

E. 結論

NDB からのデータ提供は、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集計情報も含めて、今後も量と質の増加が必要である。その方向性の一つとして、韓国 HIRA の患者サンプルデータ、米国 CMS の VRDC は今後の日本の**申出者対応部門を充実**する上で有益な先行事例である。

F.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　第 34 巻第 5 号：799-804,2016 年

G.知的所有権の取得状況

該当なし

別紙４

研究成果の刊行に関する一覧表

発表者氏名	論文タイトル名	発表誌名	巻(号)	ページ	出版年
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
松居 宏樹，大江 和彦	A Querying Method over RDF-ized Health Level Seven v2.5 Messages Using Life Science Knowledge Resources	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論文集	98-99	2015