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

諸外国の医療保険情報データベース

アメリカ、韓国、台湾におけるレセプトデータ提供の概要（厚労科研研究班調査を踏まえ作成）

	アメリカ※	韓国	台湾
レセプトデータアーカイブ機関名	Center for Medicare and Medicaid Services（CMS） メディケア等を扱う政府直轄機関	Health Insurance Review Agency 健康保険審査評価院（HIRA） 日本で言うところの審査支払機関に相当する組織であるが、一元化された組織である	National Health insurance Research Database 全民健康保険研究資料庫 日本で言うところの、国立保健医療科学院に相当する組織である
提供データの概要	CMS Identifiable Data Files: 個人を識別できる情報 Limited Data Set:個人を識別できない個票データ Non-Identifiable Data Files: 集計表 など複数	層化抽出データ 外来、入院、高齢者、小児、それぞれのタイプごとに、レセプトデータを無作為抽出して提供している このほかに、申出者の要望に応じて抽出を行いデータ提供する場合もある	系統抽出データ 日本でいうところのサンプリングデータセット 特定主題データ 疾患別に作成したデータセット など。第8回有識者会議資料（http://www.mhlw.go.jp/stf/shingi/2r9852000002d61-att/2r98520000022def.pdf）も参照
データ提供開始時期	1996年より	2012年より	2000年より
民間提供の有無	CMSが提供するデータは、データの種類により民間が利用できるものがある	HIRAの統計分析チームで公共・民間部門に提供しているが、民間提供先はマスコミや患者団体等、対象となる組織が限られている	民間への提供は、禁止している。過去に製薬会社に、治験業務サポート等の研究目的として試行的に提供した時期があったが、現在は研究者にしか提供していない
個人情報保護に関する取扱いの根拠	CMS federal Privacy Act HIPAA Privacy Rule など	個人情報には当てはまらない情報として認識されているが、個人を特定できる情報が含まれているため、実際の運用においては（個人情報に準ずる）情報として「行政機関が有する個人情報の保護に関する法律」に基づき取り扱っている	個人情報コンピュータ処理保護法（電腦處理個人資料保護法(1995年)）に則った取扱いをしている。（The Computer-Processed Personal Data Protection Law）
漏洩等、不適切利用に対する対応及びその根拠	CMS 懲役刑となる可能性もある罰則を受けることを、契約書にて承認する	申込者が利用誓約書の事項を違反した場合、あらゆる不利益および民事刑事上の処罰を甘受し、今後HIRAの資料を利用することに制限をおく	個人情報コンピュータ処理保護法（電腦處理個人資料保護法(1995年)）に罰則・罰金規定を定めるとともに、規定に違反した研究者に対しては、その研究者らに利用停止およびデータの返却を通知する。

※アメリカの事例については、第2回有識者会議での資料（http://www.mhlw.go.jp/stf/shingi/2r9852000000va02-att/2r9852000000va4b.pdf）も参照。

出典：満武巨裕 参考人資料、レセプト情報等の提供に関する有識者会議, 厚生労働省・保険局
http://www.mhlw.go.jp/file/05-Shingikai-12401000-Hokenkyoku-Soumuka/0000037075.pdf

台湾のデータ提供の事例

＜概要＞
台湾では、1995年に国民健康保険プログラムが導入され、2007年時点で総人口2,296万人中、2,260万人がプログラムに加入。このプログラムに基づく、診療報酬支払いのためのレセプトデータ等を、国民医療保険局（Bureau of National Health Insurance: BNHI）が収集しデータベースを構築しており、その管理を国家健康調査機構（National Health Research Institutes :NHRI）が行っている。
データの利用を希望する研究者は、「一般申請」と「特別申請」の2つの方法で申請ができる。

＜全民健康保險研究資料庫 (National Health Insurance Research Database)に含まれるデータ＞

レセプトファイル（原始資料檔）	備考
1．住院費用申請總表主檔（DT）	入院レセプトファイル（総表）
2．門診費用申請總表主檔（CT）	外来レセプトファイル（総表）
3．住院醫療費用清單明細檔（DD）	入院レセプトファイル
4．住院醫療費用醫令清單明細檔（DO）	入院明細（オーダーリング・処方）ファイル
5．門診處方及治療明細檔（CD）	外来レセプトファイル
6．門診處方醫令明細檔（OO）	外来明細（オーダーリング・処方）ファイル
7．特約藥局處方及調劑明細檔（GD）	調剤明細ファイル
8．特約藥局處方醫令檔（GO）	調剤レセプトファイル
9．承保資料檔（ID）	被保険者ファイル

基本ファイル（基本資料檔）	備考
1．醫事機構病床主檔（BED）	各医療機関の病床種別ファイル
2．醫事機構診療科別明細檔（DETA）	各医療機関の開設診療科ファイル
3．醫事機構基本資料（HOSB）	医療機関情報Ⅰ
4．醫事機構副檔資料（HOSX）	医療機関情報Ⅱ（開院日情報）
5．專科醫師證書主（DOC）	医師情報Ⅰ
6．醫事人員基本資料檔（PER）	医師情報Ⅱ
7．重大傷病證明明細檔（HV）	重大傷病証明明細ファイル
8．醫事機構服務項目檔（HOX）	医療機関情報Ⅲ（開院日情報）
9．藥品主檔（DRUG）	医薬品コード
10．承保資料檔（ID）	被保険者ファイル

＜一般申請＞
　あらかじめ設定された5種類のデータセットからデータの提供を受けるもの。

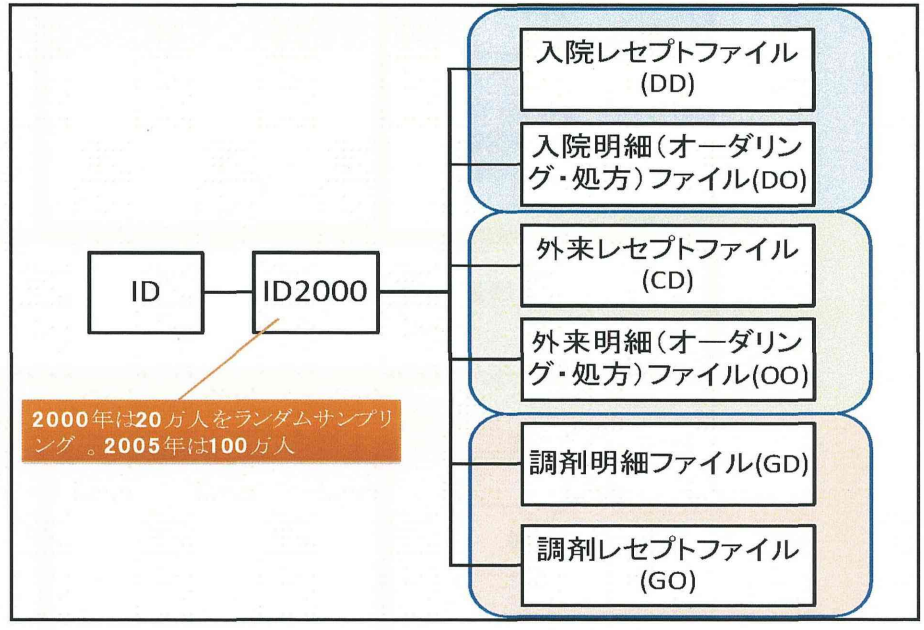
（Ⅰ）基本資料データ(基本資料檔)：集計データ
　基本ファイル（基本資料檔）の10ファイルと、DT（入院レセプトデータ（総表））とCT（外来のレセプトデータ（総表））の2ファイル。
ただし、DTとCTは医療機関単位で集計（月次）されている。

（Ⅱ）系統抽出データ(系統抽樣檔)：個票データ
　入院レセプトファイル(DD)の5%抽出
（入院明細（オーダーリング・処方）ファイル(DO)を含む）
　外来レセプトファイル(CD)の0.2%抽出(外来明細
（オーダーリング・処方）ファイル(CO)を含む）

（Ⅲ）特定主題データ(特定主題分檔)：個票データ
　特定疾患や調査目的に沿って抽出したデータファイル。悪性新生物、糖尿病、精神疾患、交通事故、リハビリ、漢方薬(中醫藥)ファイル等、現在16種類。

（Ⅳ）ランダム抽出データ(抽樣歸人檔)：個票データ
　被保険者ファイル(ID)から、ランダムサンプリングしたもの。
2000年は20万人を抽出し、（1996年～2007年の）入院、外来、調剤レセプトデータを提供している（※）。

※ランダム抽出データの2000年版ファイル構成



（Ⅴ）教育用データ(教學用資料檔)：個票データ
　2000年のレセプトデータから1,000人をランダムサンプリングしたもの。教育用に無償提供。

＜特別申請＞
　申請人が提出した研究計画に必要なデータを、当國家衛生研究院による審査終了後に、4種類のデータ（CD, OO, DD, DO）から抽出して、データセットを作成して提供する（含、調剤データ）。

（出典）：
　國家衛生研究院、National Health Insurance Research Database in Taiwanホームページ
　http://w3.nhri.org.tw/nhird/date_01.htm, http://w3.nhri.org.tw/nhird/en/Data_Files.html, http://w3.nhri.org.tw/nhird//date_03_02.php?year=97&list_n=0
　http://w3.nhri.org.tw/nhird//file_talk/workshop2008-1.pdf
　厚生労働科学研究費補助金、医療ナショナルデータベースに関する諸外国の整備状況および日本におけるデータベースのあり方研究、財）医療経済研究機構・社会保険福祉協会
（資料作成協力）：（財）医療経済研究機構・社会保険福祉協会 満武巨裕 研究部副部長

NDB研究者用データ

現在提供されているＮＤＢデータの種類

	特別抽出	サンプリングデータセット	基本データセット	集計表情報
基本的なイメージ	申出者の要望に応じ、データベースにある全データのなかから、該当する個票の情報を抽出し、提供する	探索的研究へのニーズに対応し、抽出、匿名化などを施して安全性に十分配慮した、単月分のデータセット	入院、外来、疾患別など目的に合わせて年度ごとに紐付けが可能で、簡易に分析することが可能なデータセット	申出者の要望に応じ、データを加工して作成した集計表を提供する
提供データ	個票	一部匿名化等を行った個票	大幅に加工した個票	集計表
含まれているデータ項目例	レセプト情報、特定健診等情報に含まれている、ほぼすべての項目	希少な情報があらかじめ匿名化・削除されたレセプトデータ	患者の基本属性情報以外は、主傷病名、診療識別情報、要望に応じたコードなど	集計表
利用にあたり具備すべきセキュリティ	データ利用時に、情報セキュリティマネジメントシステムを確実に運用できる利用環境を整える	特別抽出で求められるセキュリティ水準と比較して ある程度具備しやすいセキュリティ水準での利用が可能		
想定される利用者像	レセプト研究に一定の知見があり、申出内容や抽出条件を吟味し、大量のデータを高速に処理することを想定している利用者	レセプト研究に関心はあるが経験がまだ十分でなく、データの特徴や各項目の概要を把握したいと考えている利用者	レセプトの構造を踏まえながら研究するよりも、基本的項目について簡単に分析を試みたいと考えている利用者	集計された結果を必要とし、データ処理を行うことを想定していない利用者
提供実績（計66件）	42件	15件	2件	7件

2013年の提供件数は、韓国は115件、台湾は270件であり、日本の約10倍以上の実績を有する。

http://www.mhlw.go.jp/file/05-Shingikai-12401000-Hokenkyoku-Soumuka/0000117367.pdf

電子医療レセプト例

```
2,1,0,MN,910000162,東京都港区新橋,1314240591000162,,,
1,2,0,IR,1,13,1,9999913,,AAAAAA医科クリニック,42405,00,03-9999-9999
1,3,0,RE,5,1112,42404,サンプル 5,2,3450227,,,,,,,sample-ika-005,,,,,,,,,,,,,,,,,,,,,
1,4,0,HO,06132013,1234567,5,5,1130,,,,,,,,,
1,5,0,SY,0000999,4180307,1,,うつ状態,,
1,6,0,SY,0000999,4210311,1,,頚腕症候群,,
1,7,0,SY,8839792,4210515,1,,,,
1,8,0,SY,8839596,4220201,1,,,,
1,9,0,SY,7833001,4220215,1,,,,
1,10,0,SY,2809009,4220421,1,,,,
1,11,0,SY,4779004,4221008,1,,,,
1,12,0,SY,5301002,4230422,1,,,,
1,13,0,SY,3000004,4230506,1,,,,
1,14,0,SY,8844095,4230506,1,,,,
1,15,0,SY,6929365,4231203,1,,,,
1,16,0,SY,6918002,4231213,1,,,,
1,17,0,SY,7840024,4240120,1,,,,
1,18,0,SY,5319009,4240201,1,,,,
1,19,0,SY,8842865,4240221,1,,,,
1,20,0,SI,12,1,112007410,,69,4,,,,,,,,1,,,1,,,,1,1,,,,,,,,,,,,,,,,,,,
1,21,0,SI,12,1,112011010,,52,5,,,,,,,,1,,,1,,,,1,1,,,,,,1,,,,,,,,,,,,,
1,22,0,SI,12,1,112007410,,69,1,,,,,,,,,,,,,,,,,,,,,,1,,,,,,,,,,,,
1,23,0,SI,,1,112001110,,65,1,,,,,,,,,,,,,,,,,,,,,,1,,,,,,,,,,,,
1,24,0,SI,33,1,130009310,,47,5,,,,,,,,1,,,1,,,,1,1,,,,,,1,,,,,,,,,,,,,
1,25,0,IY,,1,620007328,1,,5,,,,,,,,1,,,1,,,,1,1,,,,,,1,,,,,,,,,,,,,
1,26,0,IY,,1,640454022,1,24,5,,,,,,,,1,,,1,,,,1,1,,,,,,1,,,,,,,,,,,,,
1,27,0,SI,80,1,120002710,,40,1,,,,,,,,,,,1,,,,,,,,,,,,,,,,,,,,,,,,
1,28,0,SI,80,1,120003270,,65,1,,,,,,,,,,,1,,,,,,,,,,,,,,,,,,,,,,,,
2,29,0,HO,06132013,1234567,5,5,1060,,,,,,,,
2,30,0,IY,,1,640454022,1,10,5,,,,,,,,1,,,1,,,,1,1,,,,,,1,,,,,,,,,,,,,
2,31,0,JY,2,4,0,,,29,0,
2,32,0,JY,3,25,0,,A,,,
2,33,0,JY,2,26,0,,,30,0,
2,1,0,MN,910000164,東京都港区XXXXXX,1314240591000164,,,
1,2,0,IR,1,13,1,9999913,,AAAAA医科クリニック,42405,00,03-9999-9999
1,3,0,RE,7,1112,42404,サンプル 7,2,3240506,,,,,,,sample-ika-007,,,,,,,,,,,,,,,,,,,,,
1,4,0,HO,06132013,1234567,7,4,898,,,,,,,,,
1,5,0,SY,3545003,4131225,1,,,,
1,6,0,SY,0000999,4131225,1,,不眠,,
1,7,0,SY,0000999,4140213,1,,うつ状態,,
```

電子レセプトの形式はCSV形式となっており、記号・数字とコンマの羅列である。

レコード数が患者への
診療行為(SI)、病名(SY)ごとに発生する

①診療行為マスタ:	約 6,700
②医薬品マスタ:	約 20,000
③特定器材マスタ	約 1,200
④傷病名マスタ	約 32,000
⑤修飾語マスタ	約 2,000
⑥コメントマスタ	約 300

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

3. レセプトビッグデータ解析の現状と将来

満武巨裕

健康と医療の問題はいつの時代にも大きな関心が払われてきた．少子高齢化を迎えるわが国の人口は，社会保障費用の負担が増すことが予想されるなか，医療は最も効率化が求められている分野である．近年，政府機関も文字通りビッグデータと呼ばれる膨大な量の情報を保有するようになった．その1つに，厚生労働省が2009年から収集を開始した全日本国民の医療保険データを格納するレセプト情報・特定健診等データベース（NDB）がある．本稿では，このNDBを使った分析の現状，諸外国の動向，今後の課題について解説する．

はじめに

インターネットが普及した結果として，検索ワードや購買履歴等の膨大なデータが蓄積されている．これらは「ビッグデータ」と呼ばれ，その利活用によって経済活動や社会活動に変革をもたらせると大きな期待が寄せられている[注1]．インターネットの世界だけでなく，交通，防災，エネルギー管理，医療・介護といった実世界での活動状況が現れたデータを獲得する技術（センシリング）も発達し，これらのデータに対して新たな価値づくりのための種々の処理を行うビッグデータ処理も開発されている[1]．

近年は，日本の政府機関も文字通りビッグデータと呼ばれる膨大な量の情報を保有するようになった．その1つに，厚生労働省が2009年から収集を開始した全日本国民の医療保険データを格納するレセプト情報・特定健診等データベース（以下，NDB）があり，ヘルスケア分野における最大規模のデータベースである[2]．本稿では，このNDBを使った分析の現状，諸外国の動向，今後の課題について解説する．

[キーワード]
レセプト情報・特定健診等データベース（NDB），診療報酬明細書，医療費適正化，特定健診，特定保健指導

1 レセプト情報・特定健診等データベース（NDB）

1）レセプト（診療報酬明細書）

2008年度の第5次医療制度改革の「高齢者の医療の確保に関する法律」のなかに「都道府県の医療費適正化計画の作成，検討のための資料を作成することを目的に国（担当部局：厚生労働省・保険局・保険システム高度化推進室）に必要な情報を提供しなければならない（第16条2）」，とする一文が盛り込まれたこと

注1 ビッグは主観的に「大きい」という意味であるために，ここでは定量的な基準の定義はしない．

Present and future analysis of the national medical claims database in Japan
Naohiro Mitsutake：Institute for Health Economics and Policy（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）

```
2,1,0,MN,910000162,東京都港区新橋,13142405910000162,,,
1,2,0,IR,1,13,1,9999913,,AAAAAAA 医科クリニック,42405,00,03-9999-9999
1,3,0,RE,5,1112,42404,サンプル　5,2,3450227,,,,,,sample-ika-005,,,,,,,,,,,,,,,,,,,,
1,4,0,HO,06132013,1 2 3 4 5 6 7,5,5,1130,,,,,,,
1,5,0,SY,0000999,4180307,1,,うつ状態,,
1,6,0,SY,0000999,4210311,1,,頸腕症候群,,
1,7,0,SY,8839792,4210515,1,,,,
1,8,0,SY,8839596,4220201,1,,,,
1,9,0,SY,7833001,4220215,1,,,,
1,10,0,SY,2809009,4220421,1,,,,
1,11,0,SY,4779004,4221008,1,,,,
1,12,0,SY,5301002,4230422,1,,,,
1,13,0,SY,3000004,4230506,1,,,,
1,14,0,SY,8844095,4230506,1,,,,
1,15,0,SY,6929365,4231203,1,,,,
1,16,0,SY,6918002,4231213,1,,,,
1,17,0,SY,7840024,4240120,1,,,,
1,18,0,SY,5319009,4240201,1,,,,
1,19,0,SY,8842865,4240221,1,,,,
1,20,0,SI,12,1,112007410,,69,4,,,,,,,1,,,1,,,1,1,,,,,,,,,,,,,,,,
1,21,0,SI,12,1,112011010,,52,5,,,,,,,1,,,1,,,1,1,,,,,1,,,,,,,,,,
1,22,0,SI,12,1,112007410,,69,1,,,,,,,,,,,,,,,,,,,,,1,,,,,,,,,,
1,23,0,SI,,1,112001110,,65,1,,,,,,,,,,,,,,,,,,,,,1,,,,,,,,,,
1,24,0,SI,33,1,130009310,,47.5,,,,,,,1,,,1,,,1,1,,,,1,,,,,,,,,,
1,25,0,IY,,1,620007328,1,,5,,,,,,,1,,,1,,,1,1,,,,,1,,,,,,,,,,
1,26,0,IY,,1,640454022,1,24,5,,,,,,,1,,,1,,,1,1,,,,,1,,,,,,,,,,
1,27,0,SI,80,1,120002710,,40,1,,,,,,,,,,,1,,,,,,,,,,,,,,,,,,,,
1,28,0,SI,80,1,120003270,,65,1,,,,,,,,,,,1,,,,,,,,,,,,,,,,,,,,
2,29,0,HO,06132013,1 2 3 4 5 6 7,5,5,1060,,,,,,,
2,30,0,IY,,1,640454022,1,10,5,,,,,,,1,,,1,,,1,1,,,,,1,,,,,,,,,,
2,31,0,JY,2,4,0,,,29,0,
2,32,0,JY,3,25,0,,A,,,
2,33,0,JY,2,26,0,,,30,0,
2,1,0,MN,910000164,東京都港区 XXXXXX,13142405910000164,,,
1,2,0,IR,1,13,1,9999913,,AAAAA 医科クリニック,42405,00,03-9999-9999
1,3,0,RE,7,1112,42404,サンプル　7,2,3240506,,,,,,sample-ika-007,,,,,,,,,,,,,,,,,,,,
1,4,0,HO,06132013,1 2 3 4 5 6 7,7,4,898,,,,,,,
1,5,0,SY,3545003,4131225,1,,,,
1,6,0,SY,0000999,4131225,1,,不眠,,
1,7,0,SY,0000999,4140213,1,,うつ状態,,
```

図1　電子レセプトの例

電子レセプトのデータ形式はCSV（comma separated values）となっており，記号・数字とコンマの羅列である．研究・分析に用いるためには，別途，利用者において加工する必要がある．

で，国（厚生労働省）は法的にレセプトや特定健診等のデータを収集・蓄積できるようになった．

レセプトは診療報酬請求の際に発生する業務データであり，医療保険の適用を受けている手術や注射などの約7千種類の診療行為，医薬品は約2万種類のなかから提供された保険診療行為について，患者ごとにいつ（何月），どこで（医療機関）等の情報が病名とともに記録されている．

2014年度，この保険診療行為の総額（国民医療費）が，40兆610億円まで上昇し，国内総生産（GDP）の約8.3％を占めるまでになった．国民医療費は，公的な医療保険が適用された医療費であり，日本に国民皆保険が導入された1961年から50年以上が過ぎ，国民医療費は一貫して増え続けている[3]．したがって，法律施行前（2009年4月以前）のデータがNDBに存在しないのは，残念なことである[注2]．

日本の医療制度を改革するうえで，エビデンスに基づく策定ができなかった大きな原因の1つがデータベースの不在であったため，NDBには大きな期待がかかっている．例えば，「社会保障制度改革国民会議報告書」[1]では，ICTを活用してレセプト等データを分析した疾病予防の促進，地域の将来的な医療ニーズの客観的なデータに基づく見通しを踏まえた地域医療ビジョンの策定，医療行為の費用対効果等検証のための継続的なデータ収集などのしくみの構築等の提言が行われており，医療費適正化の切り札ともいわれている．

注2　レセプトは，当初は紙であったため，電子化されるまでに大変に長い年月を要した．例えば，1983年に旧厚生省が電子化を導入しようとした際，日本医師会等の反対があり頓挫した経緯もある．

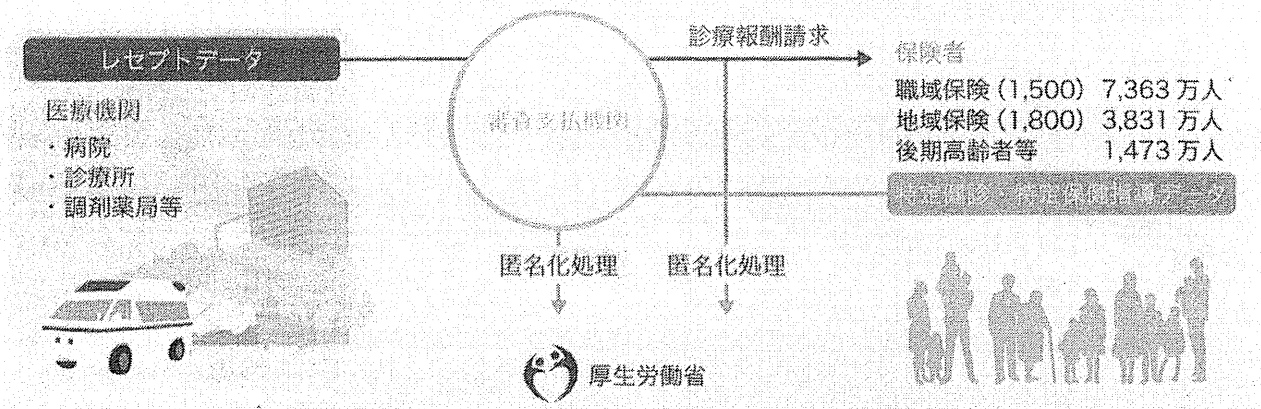


図2 レセプトと特定健診・特定保健指導データ
われわれは病気になった際，患者（被保険者）として，医療機関（診療所，病院）を受診し，診療・処置・投薬といった医療サービスを受ける．その後，医療費の自己負担分を支払う．自己負担分以外については，医療機関が月に一度，保険者に対して提供した医療保険で行われた診療行為サービスの一覧と価格を患者ごとに記載したレセプト（診療報酬明細書）を送り，請求する．

2）特定健診等データ

2008年4月に特定健診・特定保健指導制度が導入された（一般に「メタボ健診」といわれている）．メタボ健診では，40～74歳までの全国民を対象として，腹囲やBMI，血圧，問診票から喫煙に関する生活習慣，血液検査から血糖や脂質（中性脂肪およびHDLコレステロール）を測定する．例えば，腹囲が男性85 cm，女性は90 cm以上となると基準値超えとなり，加えて高血圧や糖尿病などのリスクを一定以上有する場合は，健康的な生活に改善できるように特定保健指導を受けなければならない．NDBは，このメタボ健診のデータも収集している．

現時点では，特定健診・特定保健指導を受けると医療費が低下するというエビデンスは存在しないが，NDBデータを中長期的に分析することで医療費適正（抑制）効果が得られるのではないかという期待がある．

3）データ形式と量

レセプトは業務データであり，医療機関から保険者に送付される形式は，保険診療行為の情報が羅列されたCSVファイルであり，分析しやすい形式ではない（**図1**）．NDBではこのCSV形式の電子レセプトを，複数のレコードに分割して保管している．レセプトは，主に医科（入院および入院外），DPC，調剤，歯科の4つであり年間71.7億件が発生するがさらに複数のレコードに分割され年間370億件となる．特定健診・特定保健指導データは2008年度からの約1.2億件分が蓄積されている（**図2**）．

2 NDB利用状況

1）利用者

NDBデータの利用者は，第一が厚生労働省の担当部局と都道府県であり，医療費適正化の分析やエビデンスを作成することになっている．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利用も可能となっているが，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あり，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）において承諾を得なければならない[6]．この承諾の敷居は高く，研究者等への第三者提供を検討した第1回は，43件の申出に対して承諾件数は6件であった．2012年度は9件，2013年度は3件であり，これまでの提供実績は，36件である．

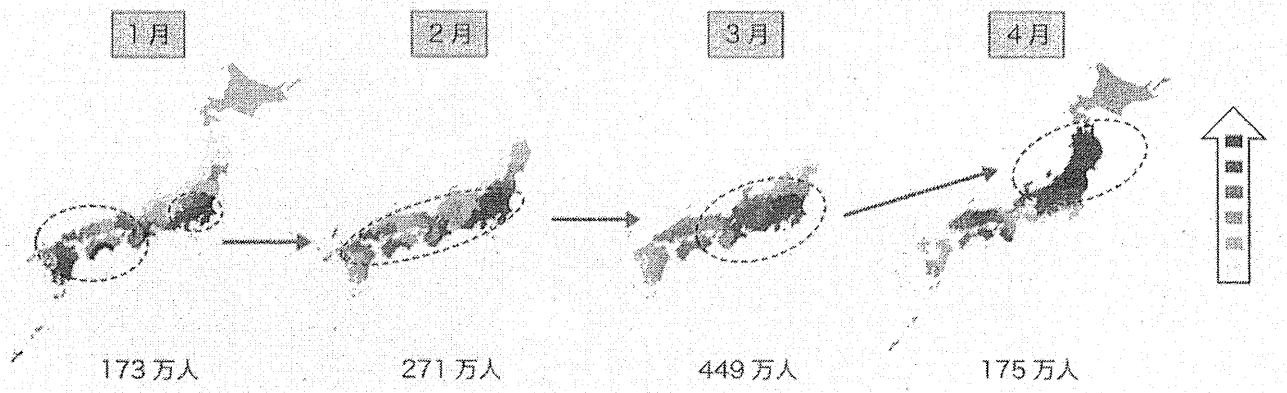


図3　アレルギー性鼻炎（外来）の月別患者数

2）少ない利用件数

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．

しかし筆者は，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の調査・分析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およ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1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を確保し，入退室記録等の装置を導入しているケースが多い．したがって，予算等の問題もあって申請を見合わせる研究者も多い．

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であるため，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すべ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いえる．

3 研究の事例

筆者は，内閣府最先端研究開発支援プログラム（FIRST）の協力を得て，NDBの2010年度の全データを扱う機会を得た[6]．利用目的が「研究用途における汎用性の高いレセプト基本データセットの設計と作成を行う」ことであったために，自由な分析はできなかったが，いくつかの成果を得ることができた[7]．

これまで日本における患者数については「患者調査（厚生労働省）」が該当の基幹統計であるが，“入院及び外来患者については，10月中旬の3日間のうち医療施設ごとに定める1日”の調査データをもとに推計を行っているサンプリング調査であるために，例えば冬の時期に流行するインフルエンザや春先の花粉症などの患者は測定できなかった．

一方，全レセプトデータを有するNDBは，年単位および月単位で患者数を集計できる．例えば，アレルギー性鼻炎（入院外）での患者数は，1月に173万人が発生しており，2月は271万人，3月は449万人とピークを迎える．しかし，4月には175万人にまで落ち込む．また，都道府県でみると九州から関東，東北に疾患の発生状況が移動しているのが見てとれる（**図3**）．

2012年度のデータから，電子レセプトに日計表が義務付けられた．したがって，今後は月ごとよりも細かい日々の患者の発生状況や投薬実態についての分析が進むであろう．また，NDBは，病名だけではなく，診療行為データを含むため，例えばインフルエンザ治療

薬がどの時期にどの地域の医療機関でどれだけの量が処方されたか（使用量）も判別可能である．

おわりに：NDBの課題

NDBの利用が進むにつれて，保有するデータの精度についても検証の必要性があることがわかってきた（これまで，全NDBデータの精度検証は公表されていない）．

筆者は2010年度のデータを使って，以下の3点を公表している．第一に，診療所の入院外レセプトと突合できない調剤レセプトが存在するため，外来の医療費が実際よりも少なく推計される可能性があること．第二に，特定健診データとレセプトデータを突合できない保険者が存在する（レセプトと特定健診データに関するリンケージ率の低さを指摘した論文は近年公表された）．第三として，NDBのIDには欠点があり，ユニークな番号が日本国民の人数を超えてしまう点などがある．

さらに，レセプトだけでは，死亡情報が正確に把握できない．この解決策としては，各保険者が保険料の徴収や加入者の確認のために日々更新している被保険者台帳といわれるマスタを収集すればよい．被保険者台帳には，保険者の異動（例えば透析を受けることになり生活保護へ保険者が変更になった）などの情報も含まれている．

特定健診・特定保健指導データについても同様の課題があり，受診した被保険者のみのデータしか収集してしないため，未受診者の分析ができない．特定健診・特定保健指導の対象となった集団の被保険者台帳も各保険者が保有しているので，レセプトと同様の改善が期待できる．

上記にあげたような課題が存在するものの，NDBは世界でも類のない貴重なデータベースである．その理由としては，日本は国民皆保険制度が導入されているために，日本全国の医療機関（病院，診療所，院外薬局等）で行われた保険診療行為の記録がすべて取り込まれている．加えて，少子高齢化が進み人口減が予測されるものの，現時点では世界で10番目である日本の人口の悉皆ビッグデータであることがあげられる．

日本と同様の皆保険制度であり診療報酬点数制度を導入している国として，韓国と台湾がある．両国は，日本の利点と欠点を十分に調査したうえで国民皆保険制度を導入したため，レセプトの電算化も同時に実現している．データの研究利用も盛んであり，研究者へのレセプトデータ提供件数は年間100件を超えている（2013年の韓国の研究者へのレセプトデータの提供は115件，台湾は270件）．韓国および台湾のレセプトデータ研究利用申請者は，誓約書を提出し研究承諾が得られれば，日本のような利用規定に縛られることなく分析が行える．日本のNDBデータは，個人情報が除去されており，データ利用がはじまりすでに4年が経過しているが規定違反や重大なアクシデントも発生していないことから，利用規定を緩和する時期に来ていると思われる．

日本の医療業界のビッグデータの活用はまだまだ発展途上の段階だが，近年は，NDBよりもはるかに大量のデータのデータベース化と，これらのデータを組合わせて高速処理するデータベース技術が開発されている．

健康と医療の問題はいつの時代にも大きな関心が払われてきており，少子高齢化を迎えるわが国の人口は，2030年には約1億人に減少し，約40％が65歳以上の高齢化社会となり，医療費に加え介護費や年金を含む社会保障費用の負担が増すことが予想される．そのために，医療は最も効率化が求められているといえ，さまざまな精巧な予測のモデルとビッグデータ処理技術が融合することで，医療費適正化の有益なツールとなることが期待されている．

文献

1）「角川インターネット講座7 ビッグデータを開拓せよ 解析が生む新しい価値」（坂内正夫／監修），角川学芸出版，2015
2）満武巨裕：日本のレセプト情報・特定健診情報等データベース（NDB）の有効活用．情報処理，56：140-144，2015
3）「平成25年度国民医療費」，厚生労働省大臣官房統計情報部，2016
4）首相官邸 社会保障制度改革国民会議報告書，2013 https://www.kantei.go.jp/jp/singi/kokuminkaigi/pdf/houkokusyo.pdf
5）厚生労働省 レセプト情報・特定健診等情報提供に関するホームページ http://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryouhoken/reseputo/info.html
6）内閣府最先端研究開発支援プログラム（FIRST）「超巨大データベース時代に向けた最高速データベースエンジンの開

発と当該エンジンを核とする戦略的社会サービスの実証・評価」(FIRST中心研究者：喜連川 優)
7) 平成24年度～平成25年度厚生労働科学研究「汎用性の高いレセプト基本データセット作成に関する研究」(研究代表者：満武巨裕)

<著者プロフィール>

満武巨裕：2004年，京都大学大学院人間・環境学研究科博士後期課程単位取得退学．1998年，米国・スタンフォード大学アジア太平洋研究センター客員研究員．'05年，東京大学医学部附属病院22世紀医療センター健診情報学講座研究員．'06年，財団法人医療経済研究機構主席研究員/副部長（現在に至る）．'15年から厚生労働科学研究事業・戦略研究の研究代表者として，「レセプト情報・特定健診等情報データベースを利用した医療需要の把握・整理・予測分析および超高速レセプトビックデータ解析基盤の整備」に従事している．

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～

松居 宏樹[*1] 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・経済学分野
[*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

Usage of the Japanese national insurance claims database in onsite research center. ~Operability, usability and limitation~

Matsui Hiroki[*1] 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 an on-site research center, which enables health care researchers to analyse Japanese administrative claim database (The Japanese National Insurance Claims Database: NDB). However, difficulties in research implementation and computer system workload have not been investigated. In this presentation, we show an overview of our computer system workload investigation and discuss about difficulties in research. Most of previous studies with NDB have shown descriptive statistics from cross-sectional or time-series cross-sectional data as a final result. On the other hand, researches using cohort/panel data enforce researchers complex data-handling process. Researcher friendly standardized data formats and sharing data-handling script could make researchers to avoid complex data-handling process. We had been creating data-handling script which extracts and formats target patients data. Although avoiding complex data-handling process, software restriction and system storage size restriction are still large difficulties in research implementations. Further efforts are needed for extending NDB research use.
Keywords: Administrative claim database, Data handling, Epidemiology

Keywords: Administrative claim database

1. はじめに

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からは医科・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る．これにより，研究者が自前で解析環境やデータ保管環境を用意する事を回避できることから，今後レセプト情報等はより研究に利用しやすくなる物と考えられる．

しかし，当該システムの機能で保健医療にかかわる研究者が実際に行う研究が滞りなく実施できるか，システムの性能は十分か，研究を実施する上で研究者にはどのような技能・技術が必要か，またそれは利用普及の障壁となりうるか，今後どういった点を改善すべきか，などはいまだ明らかではない．

また，複雑なデータ切り出しや高度な統計分析がシステム上可能であったとしても，その負荷によっては，一般利用者が一斉に利用した場合システムに過剰な負荷をかけることになりかねず，負荷量に応じて事前にタイムスロットを割り当てるなどのマネジメントが必要となるため，さまざまなケースを考えた負荷量の想定が試行運用期間中に把握される必要がある．

東京大学では，オンサイトセンターを稼働させるにあたり，大学院医学系研究科公共健康医学専攻の4講座が連携しオンサイトセンターのパフォーマンスを試験するパフォーマンス研究の申請を厚生労働省に行い，試験の準備を進めている．ここでは，パフォーマンス研究の過程で明らかとなりそうな，レセプト情報等オンサイトリサーチセンターの課題と，それに対する我々の取り組みを紹介する．

2. レセプト情報の研究利用とその課題

過去，厚生労働省が提供を行ったレセプト情報を用いた研究の多くでは，レセプト情報を医薬品や診療行為の実態調査や疾病の発生・有病率の推計，医療費の計算などに用いてきた．また，地域別の医療受給の推計や，健診データを用いた解析なども行われてきた．

これらの研究の多くは，クロスセクショナルもしくは，多時点クロスセクショナルデータを用いた記述統計が最終的な成果物となってきた．オンサイトセンターでは，システムに導入されたOracle Business Intelligence (BI ツール) を用いて，記述統計を行う事が可能であることから，今までに多く行われてきた研究と同様の解析を比較的容易に実施できると考えられる．しかし，研究者が個票レベルのクロスセクショナルデータを用いた多変量解析等を行おうとする場合，データをBIツールで出力するには大きな困難が伴う．BIツールがそもそもデータを抽出・整形する事よりもデータの迅速な集計を目的として作成されていること，BIツールが出力で

きるテーブルの行数に制限があること等が障壁となっている.そのため,レセプト情報等を充分に研究利用するためには,研究者自らが工夫してデータを抽出・整形する必要がある.

また,国外の事例では,大規模な医療の請求データを用い,個人や医療機関・地域を経時的に観察したコホート・パネルデザインでの医療技術評価や,保険医療政策の評価,疫学研究が行われてきた.オンサイトリサーチセンターに導入されたBI ツールには,コホート・パネルデータを出力する機能が備わっていないため,これらの研究デザインを用いた研究を行う場合にも,研究者は工夫してデータを抽出・整形する必要がある.

このように,レセプト情報を用い,記述統計に留まらない様々な研究デザインを遂行する上でデータのハンドリングを研究者自らが行う必要がある.これは,リレーショナルデータベース管理システムにある程度の知識を有し,SQL などの問合せ言語を用いて大規模データを扱った経験を有する研究者には,それほどハードルが高い課題ではない.しかし,保健医療にかかわる研究者の全てがそれらの知識と技術を充分に有しているわけではない.つまり,オンサイトセンターを保健医療にかかわる研究者が利用する上で,データハンドリングのコストが高い事が大きな課題となっている.

3. オンサイトセンター若手実務者会の取組

データハンドリングのコストを下げるためには,保健医療研究者が利用する標準的なデータ形式を設定し,そのデータ形式でデータを出力するプログラムを作成し研究者間で共有することが効率的である.そこで,東京大学では,医療政策・医療情報・臨床疫学等の研究を行い,リレーショナルデータベース管理システムシステムの知識とSQL などの問合せ言語を作成する技術を有し,大規模データを扱った経験を有する若手研究者で実務者会を組織した.その上で,オンサイトセンター利用を見据えて,自らが必要とする情報を含んだ標準的なデータ形式の設定や,コホート・パネルデザインに適したデータ抽出・整形フローの設計,それらを実際に行う抽出プログラムの作成を行ってきた.

オンサイトセンターの環境と実際のプログラミング

オンサイトセンターにはNDB を操作するためのインターフェースとして BIツール とOracle R Enterprise がインストールされている.また,NDB を直接閲覧することは出来ないものの,SAS が環境下にインストールされた端末がある.我々はOracle R Enterprise を介して稼働するR のプログラムとして上記のシステムの構築を進めてきた.

データ形式の設定

保健医療研究者がレセプトデータを研究利用する上で,必要となるデータ粒度を「患者個人レベルのデータ」「患者エピソードレベルのデータ」「患者実施日レベルのデータ」に整理し,それぞれをテーブルとして出力した.研究者はこれらのテーブルを突合することにより,統計解析に投入するデータを作成することが可能で有り,研究者のデータハンドリングコストを大きく下げることが出来ると考えられる.

患者個人レベルの情報

患者の個人ID単位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,観察終了月での死亡有無をカラムとして含む.

患者エピソードレベルの情報

患者の外来受診や入院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.

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をカラムとして含む.平成24年度以降データについて作成する.

データ抽出とコホート・パネル化

個人を追跡して観察するコホート・パネル形式のデータを作成するためには,個人ID が同一であるレセプト情報を全て抽出し,それらを上記で定義したデータ形式に整形する必要がある.NDB には個人を識別するID として保険者ID,被保険者票番号,生年月日,性別を元に作成されたハッシュ値(ID1),氏名,生年月日,性別を元に作成されたハッシュ値(ID2) が用意されている.ID1とID2を組み合わせた個人追跡テーブルを生成することでデータのコホート・パネル化に対応した.

4. 残されている課題

上記の様にデータハンドリングにかかるコストを低減したとしても,研究の障壁になる課題は残されている.オンサイトセンターのシステムとしての課題としては,システムの保守の関係で使用できるソフトウェアが限定されている事や,ハンドリング出来るデータサイズに限界がある事があげられる.また,今後利用者が増えた際にどの程度運用面で負荷が生じるかは不明であるため,今後継続的に分析と改善を行っていく必要がある.また,利用者に求められる最低限の情報処理技術についても現状要件はなく,今後教育や情報提供を合わせて検討をして行く必要性がある.

また,オンサイトセンターのシステムとしての課題に加えNDB のデータベースの特性としての課題が存在している.これらの課題についてはすでに多く言及がなされているが,疫学研究などを遂行する上では,アウトカムが限定的であり死亡の有無が把握しにくい点や,個人の識別・追跡が不完全なIDによってしか行えない点が挙げられる.また,重要な患者重症度がほとんど取得できないうえ,現状は病院単位であっても他のデータベースとの突合が出来ない事から極めて限定的な情報しか解析に用いることが出来ない点は大きな限界である.これらに対して,技術面・制度面の両面から対応していく事が今後求められてくると考えられる.