20150102-2A

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に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

基本データセットの利活用に関する検討

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

研究要旨

本研究は、レセプト情報・特定健診等情報データベースの利活用を促進するために、厚生労働省および有識者会議の許可を得た「基本データセット」について、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を行った。

①抽出条件を脳神経外科の専門医の知見を得て検討し、抽出プログラムを作成した。

②データベースからデータセットを生成する抽出プログラムを作成した。

③データ精度および基本統計量を確認し、基本データセットの利点と課題を明らかにし、解決に向けた提案を行った。

本研究の結果から、基本データセットは分析しやすいレコードフォーマットの形で利用することができるが、抽出精度を高めるためのさらなる検討や、基本データセットの制約条件についての見直しが必要であることが明らかとなった。

A.研究目的

本研究は、レセプト情報・特定健診等情報データベースの利活用を促進するために、新たなサンプリングデータとして申出可能となった「基本データセット」について検討を行った。基本データセットとは厚生労働省・有識者会議で許可されたサンプリング抽出条件下での患者毎・レセプト毎のパネルデータであり、匿名化等のセキュリティ対策を施した汎用性が高く分析が容易かつ、利用に係る手続き及びセキュリティ要件がサンプリングデータセットの手続と同等のデータセットである。

本研究は基本データセットを学術・政策研究に資するデータとして利活用するために、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を対象に、以下の観点から検討を行った。

1．抽出項目の設定方法
2．データセットを生成する抽出プログラム
3．抽出したデータセットの精度・基本統計量

これらの検討により明らかとなった基本データセットの利点と課題を整理し、解決に向けた提案を行った。

B.研究方法

基本データセットは 2009 年度、2010 年度、2011 年度の全レセプト・特定健診・特定保健指導データから、都道府県での区分が可能な保険者（協会けんぽ、市町村国保等）から 47 都道府県を 8 ブロックに分け、それぞれのブロックから都道府県をランダムに人口比約 25%となるよう

に選定している。保険者は 20%をランダムに選定し抽出し、医療機関コードおよび保険者番号は全く異なる一意の番号を付与している。

基本データセットは、DPC レセプトおよび入院レセプトから以下のレコードフォーマットで構成される。

【データセット C】: 保険者番号・被保険者証・生年月日・性別から生成される ID1、氏名・生年月日・性別から生成される ID2、男女区分、年齢階級（5 歳区分）、入院日、保険種別、合計点数、合計診療実日数（入院）、医療機関コード（匿名化）

【データセット D】: データセット C に加え、社会保険表章用疾病分類[1]（以下、「121 分類」という）、主病名、主病名フラグのある傷病名コード（最大 5 つ）、主病名フラグのない傷病名コード（最大 5 つ）、その他申出者が自由に設定した診療行為・医薬品等のフラグ（最大 256 項目）

これらのデータセット構造に基づき、検討を行った。

1．抽出項目の設定方法の検討

（1）抽出対象は以下の通り設定した。

・レセプト種別は DPC レセプト、医科レセプト（入院）を用いた。

・傷病名（121 分類）は b-0906（脳梗塞、脳梗塞の続発・後遺症）、b-0908（その他の脳血管疾患、非外傷性硬膜下出血、脳卒中、脳血栓症）とした。

・抽出期間は入院日が平成 22 年度の入院患者レセプトとした。

・平成 22 年度に退院しなかった患者のデータは打ち切りデータとした。

・入院日が平成 22 年度以前であり、平成 22 年度も入院中の患者は除外した。

・抽出項目は、「社会保険診療報酬支払基金レセプト電算処理システムマスターファイル」に基づき、平成 22 年度の傷病名および診療行為ならびに医薬品コードを指定した。

（2）データセット C において、基本データセットの個人情報特定リスクに対して慎重な判断が求められたことから、診療開始日、入院回数、転帰区分は抽出されないこととなった。

（3）データセット D の「申出者が自由に設定できる診療行為等」については、脳神経外科の専門医の知見を得て、関連する入院基本料の他、手術、処置、リハビリテーション、医薬品等から該当するレセプト電算コード、全 242 項目を指定した。基本データセットの抽出項目数の上限は傷病名や診療行為の電子レセプトコードを単位として 256 項目のため、抽出項目が大幅に制限された。

たとえば脳梗塞の急性期治療に用いる代表的な医薬品である「エダラボン点滴静注」の場合、数量単位や製薬メーカーの違いにより 41 種類のコードが存在する。その結果、医薬品の項目については抜本的な見直しを行い、合併症に係る医薬品や使用頻度の低い医薬品については抽出対象から除外せざるを得ず、医薬品

---

[1]「社会保険表章用疾病分類」とは、世界保健機関（WHO）より公表されている「疾病及び関連保険問題の国際統計分類」（略称、国際疾病分類：ICD）に準じて定められたものであり、社会保険の分野で疾病統計を作成する際の統一的基準として広く用いられている（厚生労働省ホームページより引用）
http://www.mhlw.go.jp/bunya/iryouhoken/database/zenpan/shobyo_bunrui.html

コードの総数は86項目に限定した。そのため高血圧症や糖尿病等の合併症に係る診療行為は抽出項目から除外した。

（4）DPCレセプトの場合、医薬品は「出来高算定」と「包括算定」のいずれかによって抽出するレコードファイルが異なる。前者はIYレコード、後者はCDレコードである。また、後発医薬品については採用年度によって抽出されないため、医薬品マスタの採用年月に留意した設定が必要である。

2．データセットを生成するプログラム

（1）基本データセット用のレセプト情報等データベースの傷病名データは、一つの項目に複数の電子レセプトコードが縦棒で区切られた要素で格納されている。このデータベースから指定した電子レセプトコードを含むレセプトを抽出する場合、SQLによるクエリやスプレッドシートのフィルタ機能等で抽出することができないため、プログラムが別途必要であることが明らかとなった。

（2）基本データセットは、指定した診療行為の電子レセプトコードがあれば出力フラグ1を入力する形式で提供される。そのため、指定した電子レセプトコードをそれぞれ検索しフラグを入力する抽出プログラムを作成した。

（3）データセットは一連の入院を1つの「エピソード」として出力する。そのため出力された1行は、入院期間が複数月であっても1入院となる。ただし転院や再入院の場合は複数行に出力される。

3．データセットの精度・基本統計量

抽出したデータセットのレコードフォーマットを確認し、レコードフォーマットの構成とデータ精度を確認し、基本統計量を算出した。

C.研究結果

1．抽出結果概要

（1）抽出した基本データセットのデータ件数は101,423件の入院エピソードと292,745件のレセプトとなった。抽出項目の設定にあたっては、脳神経外科医の専門的知見により、脳血管疾患にかかわる主病名を確認するとともに、急性期治療で用いる代表的な医薬品についてガイドラインや文献に基づき指定した。

（2）121分類で指定した傷病名は、主病名フラグのある傷病名コードと、主病名フラグのない傷病名コードからそれぞれ抽出した。

2．抽出プログラム

（1）プログラム処理を行うため、指定した電子レセプトコードを1列目のみで構成する複数行ファイルを、「診療行為名」、「傷病名」、「医薬品名」の3つに分類しそれぞれ作成した。

（2）データベースに格納されている全データのうち、上記3つのファイルで指定した電子レセプトコードを1つでも一致するコードがあれば列の末尾に出力フラグ1または0を入力する列を追加し、これを繰り返す処理を行い、出力フラグ1に該当するレセプトのみを標準出力するプログラムを作成した。プログラムの実行はデータベース管理者に依頼した。

（3）抽出されたデータセットの確認を

行い、厚生労働省・有識者会議が定めるサンプリングデータセットの申出手続きに従い、基本データセットの提供を得た。

3．データセットの精度・基本統計量

患者属性

性別は男性が 53.5%とやや多く、年齢区分は 75 歳以上が 75.8%であった。

傷病名構成については、主病名フラグのある傷病名コード（最大 5 つ）の先頭に記載された第 1 主病名は「脳梗塞(I63.x)」が 54.4%、「脳血管疾患の続発・後遺症(I69.x)」が 9.1%であった。第 2 主病名、第 3 主病名、第 4 主病名になるに従い、これらの割合は減少し傷病名の種類数が増加する傾向が見られた。第 5 主病名は電算コードの桁数が異なっていたため、集計が不可能であった。

入院基本料種別

一般病棟の入院患者が多く、特定機能病院の入院患者は少ない結果となった。

また、121 分類別の入院日数および診療点数の平均を比較した結果、高齢ほど入院日数は長く診療点数は低かった。一般病棟と特定機能病院は、入院基本料が高い病棟ほど診療点数が高い結果となった。療養病棟は一般病棟と比べ診療日数が長いが、診療点数は低い結果となった。

診療行為の実施割合

平成 22 年度に入院した患者のみを抽出したにもかかわらず、脳梗塞の急性期治療を行ったレセプト件数は手術 1.5%、処置 19.6%、検査 19.4%、医薬品 10.8%と総じて低い結果となった。脳梗塞に関連するリハビリテーションを実施したレセプトは 67.9%であった。

診療行為の内訳については、手術は動脈血栓内膜摘出術（0.19%）、脳血管内手術(0.14%)、医薬品はラジカットが 7.4%、ヘパリンナトリウム注が 2.18%であった。いずれの診療行為も脳梗塞の急性期治療として必須であることから、実際に脳梗塞を発症した患者の抽出精度は低い結果となった。

D.考察

本研究は、レセプト情報・特定健診等情報データベースの利活用を促進するために、新たなサンプリングデータとして申出可能となった「基本データセット」について検討を行った。

1．抽出項目の設定方法に関する考察

（1）基本データセットの利点は、サンプリングデータセットが 10 月診療分の単月データのみの提供であるのに対し、基本データセットは最大 3 年間のパネルデータを入院エピソード単位で利用できる。また、サンプリングデータセットは出現頻度の低い傷病名や診療行為等の匿名化や高額なレセプトは削除されている等、抽出項目に厳しい制限があるが、基本データセットは、申出者が自由に設定した診療行為や医薬品等を最大 256 項目まで設定することができる。また特別抽出と異なり、分析しやすいレコードフォーマットの形で提供を受けることができ、高度なデータ処理プログラムの知識や経験を有しない研究者も利用することができる。

（2）一方、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。抽出項目を限定すること作業は、臨床の専門的知見を以てしても困難であるだけでなく、高血圧や糖尿病等のように医薬品の種類が多い場合、診療情報が極めて制限される。この制約条件により、特定の疾患等、抽出条件によってはレセプト情報等を用いた学術研究に資するデータの要件を満たさない可能性があることが懸念される。

2．抽出プログラムの作成

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。今後、申出者が継続的に利用可能な体制を整えるためには、抽出プログラムの汎用化やデータ仕様の提示等、申出に対するデータ抽出体制の強化について検討が必要不可欠である。

3．データセットの精度・基本統計量

本研究で抽出した基本データセットは平成 22 年度に脳梗塞を発症したレセプトを対象とした。しかしながら急性期治療を行ったレセプト件数は数%程度であり、抽出精度は低かった。また、本研究の結果は、厚生労働省の調査による脳梗塞の年間患者数117万9000人と大きく乖離していた。その原因として、平成22年度以前に入院したレセプトを除外していることや一部の保険者、医療機関、都道府県を用いたために偏りが生じたことが考えられる。加えて、脳梗塞の発症時期が不明のため既往疾患の患者が混在して抽出された可能性も考えられる。

また、保険者や都道府県を限定したことで実際に急性期治療を行った件数が数百件程度となった結果、患者構成に偏りが生じた可能性がある。これらの項目は匿名化されているため、データの精度や学術的利活用の観点から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。

E.結論

本研究は「基本データセット」の概要および実際の利用方法やデータの質について検討を行い、基本データセットの利点と課題を整理することができた。

本研究の結果から、基本データセットは分析しやすいレコードフォーマットの形で利用することができるが、抽出精度を高めるためのさらなる検討や、基本データセットの制約条件についての見直しが必要であることが示唆された。

また、基本データセットの利用を促進するためには、申出者が利用可能な抽出プログラムの作成や抽出作業の標準化等、管理運営体制等を整備する必要があることが示唆された。

参考文献

1) 厚生労働省第 17 回レセプト情報等の提供に関する有識者会議　資料 3「基本データセットについて」
2) 厚生労働省第 18 回レセプト情報等の提供に関する有識者会議　資料 6「基本データセットにおける匿名化基準等について」
3) 厚生労働省第 29 回レセプト情報等の提供に関する有識者会議 資料 2「基本データセットの提供について」

4) 社会保険診療報酬支払基金レセプト電算処理システムマスターファイル http://www.ssk.or.jp/seikyushiharai/tensuhyo/kihonmasta/index.html
5) 加藤 源太、平野 景子、赤羽根 直樹「レセプト情報・特定健診等情報データベースの利活用について」（統計,65,10,8-13）

F.健康危険情報
該当なし

G.研究発表
該当なし

H.知的財産の出願・登録状況
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

ataindex	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