令和 5 年度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））
臨床疫学に活用可能なNDB等データセットの作成に関する研究（21AA2006）

分担研究報告書

HIC における分析用データベース構築における課題検討

研究代表者　森　由希子　京都大学医学部附属病院医療情報企画部　講師
研究分担者　加藤　源太　京都大学医学部附属病院診療報酬センター 特定教授
研究分担者　黒田　知宏　京都大学医学部附属病院医療情報企画部　教授
研究分担者　植嶋　大晃　京都大学国際高等教育院附属データ科学イノベーション教育研究センター　特定講師
研究分担者　大寺　祥佑　国立研究開発法人 国立長寿医療研究センター 老年学・社会科学研究センター 医療経済研究部 副部長
研究分担者　田宮菜奈子　筑波大学医学医療系/ヘルスサービス開発研究センター 教授
研究分担者　杉山　雄大　国立国際医療研究センター 医療政策研究室長
研究協力者　小宮山　潤　筑波大学医学医療系　助教

研究要旨
NDB は令和 2 年 10 月から介護 DB 及び令和 4 年 3 月から DPC データベースとの連結が可能となった。さらに「医療・介護データ等の解析基盤（Healthcare Intelligence Cloud : HIC）」がクラウド環境で構築・運用開始されており、今後利用者の利便性が向上することが期待される。本研究では HIC の試行利用を行い、HIC の利用における課題抽出および解決法の検討を行った結果を報告する。

A．研究目的

レセプト情報等を活用した分析の体制整備の推進および保健医療に関するビッグデータの効果的な利活用を推進するため、NDB や、NDB と介護 DB の情報の連結解析を可能とするために医療介護連結解析基盤（Healthcare Intelligence Cloud : HIC）が構築された。現在 NDB データの提供方法として

・特別抽出（審議会による提供承諾を得た研究において、研究者の依頼に応じたデータ抽出を行い、個人単位のデータを提供）
・集計表（審議会による提供承諾を得た研究において、研究者の依頼に応じたデータ抽出を行い、作表して提供）
・サンプリングデータセット（トライアルデータセット）（1 カ月分の匿名レセプトデータに対し抽出・匿名化処理等を行ったレディメイドのデータを提供）
・オンサイトリサーチセンター（厚生労働省が指定する情報セキュリティ対策が講じられた施設において、厚生労働省が管理す

る NDB と通信回線で結ばれた端末の利用環境）
がある。オンサイトリサーチセンターでのデータ利用以外では、利用者は研究計画に基づいて抽出されたデータの提供を受け、自身で準備した分析環境において解析を行う。
HIC はオンサイトリサーチセンターと同様、クラウド上に設置された NDB データの分析環境であるが、従来のオンサイトリサーチセンターでのデータ利用と異なり、申請に基づき抽出されたデータが HIC 上に提供され、利用者はすでに抽出されたデータを HIC 上で分析するため、いわゆる特別抽出データやサンプリングデータセットの利用環境としての側面が強い。今回特別抽出データを HIC 上で利用するケースにおける課題及びその解決法について検討した。

Ｂ．研究方法

利用承諾を受けた NDB・介護 DB 連結データを用いて、HIC 上での分析環境構築し、以下の課題について検討を行った。
・データ格納
・windows 環境における分析環境構築
・Linux 環境における分析環境構築

今回利用したデータは以下の通り。
・NDB データ（医科入院・医科外来・DPC・歯科・調剤）4 年分（HIC 利用開始時は 3 年分）
・介護 DB データ　3 年分
（HIC 利用開始時のデータ量は全 csv ファイルで約 11TB）

Ｃ．研究結果

（1）データ格納

HIC は AWS Cloud 上に環境が構築されており、提供されるデータは s3 領域に格納される。利用者は s3 からデータを EC2 instance にダウンロードし、EC2 instance に与えられた領域で分析を行う。今回 HIC 利用開始時に提供を受けたデータは全 csv ファイルで約 11TB。試行利用において、EC2 instance に与えられたストレージは約 16TB であった。

（2）EC2 instance におけるデータベース構築（windows 環境）

EC2 instance に以下の手順で PostgreSQL データベース作成を試みた。

① NDB　DWH を作成
② 介護 DB DWH を作成
③ ①と②を名寄せしデータセットを作成

しかし、Postgres に提供データを登録した時点でデータ量が約 1.2 倍増加し、さらに名寄せ情報付与用の領域が必要であり、この時点で研究者作業用の領域が確保できないことが判明した。このため、ストレージ容量不足に対して以下の検証を行った。（図 1）

(ア) 対受領データに対するアプローチ
【不要カラムの削除の検証】
・提供不可カラムの削除　→効果なし
・NULL のカラムの削除　→効果なし
・不要となる ID 系カラムの削除→1,434GB（約 11%）
【データの型変更の検証】
・ハッシュ値を数字で振りなおす→174GB（約 1%）
・フラグ系カラムの最適化→848GB

（約 7%）
・年月系カラムの最適化→903GB（約 8%）
・数値系カラムの最適化→51GB（約 0.5%）
検証結果として全約 12TB のデータに対して total で約 2.4-.3.4TB の削除が可能であったが、十分な作業領域の確保は困難であった。

(イ) 対利用環境に対するアプローチ

【PostgreSQL 圧縮機能の利用の調査】
PostgreSQL におけるデータ圧縮機能である TOAST 圧縮（PGLZ、lz4）の活用を検討したが、適用可否の判断が 1 レコードの長さに依存するため、NDB データは適用対象とならず。対応不可であった。

【s3 をネットワークドライブ利用する方法の調査】
ネットワークドライブのように、s3 を”マウント”するアプリは存在したが、PostgreSQL などでファイルを参照する際、”s3 からダウンロードして開く”という挙動のため、ストレージ容量不足に対しては 効果がなかった。

【s3 上のファイルを外部参照する方法の調査】
PostgreSQL のアドオンツール（s3csv_fdw（フリーソフト））で、s3 上の CSV ファイルをテーブルとして参照する機能が存在したが、Windows 非対応のため、Windows 環境では利用できなかった。

（3）EC2 instance におけるデータベース構築（Linux 環境）

【s3 に対する SQL 実行（S3_fdw）】
Postgres の拡張機能である Foreign Data Wrapper (fdw) を利用し、s3 に直接アクセス可能かを検証したが、権限なしエラーが発生し実行できなかった。

【PostgreSQL によるデータ圧縮（citus）】
Postgres の拡張機能である Citus を用いて、実現性と圧縮可否について検証を行った。テストデータにおいて、圧縮有と無のテーブルを比較し、問題なく圧縮されていることを確認した後、実際のデータを用いて圧縮率の検証を行った。Citus を用いて 4 年分の NDB について圧縮を行ったところ、全体で約 20TB のデータを Postgres 上において約 4TB までの圧縮が可能になることが分かった。この作業により、分析用の領域の確保が可能になった。一方で、INDEX については圧縮を行うことができないため、複数の利用者が同一領域を利用する際には特に注意が必要なことも分かった。

D. 考察

【ストレージの課題について】
今回 HIC の試行利用にあたり、当初 Windows 環境での分析環境構築を試みたが、ストレージの問題から EC2 instance における PostgreSQL データベース作成ができなかった。このため容量不足に対するいくつかの検討を実施したが、Windows 環境下では分析に十分な容量の確保はできなかった。次いで Linux 環境での検討を行った。Citus を利用することで、PostgreSQL のデータを圧縮することが可能となり、一定の分析用領域の確保が可能となったが、実際に分析を行う場合には、INDEX については圧縮を行うことができないことなど、圧縮後も容量超過に対していくつかの注意が必要なことが分かった。今回は試行利用としてストレージについても通常より多い 16TB の

環境において検討を行ったが、データベース構築および分析用の領域の確保には様々な工夫を要した。実際に HIC で分析を行う際は提供されるストレージの容量内で分析可能なデータサイズを認識しておく必要がある。また、今回のように複数チームで領域をシェアすることや、解析中に発生する中間生成物の規模を考えると、ストレージについてはある程度余裕があることが望ましい。

【解析ソフトウェア環境について】

今回、データベース構築に時間を要し、実際のデータ分析は十分に行えなかったが、今後実際に分析を行うことを想定すると、解析ソフトウェア環境についても、singularity 等のコンテナが使える環境が望ましいと考える。（R のパッケージのインストールについてはパッケージ間の依存関係があるため）

【利用者への情報提供について】

今回の試行利用を行った経験から、現状において、このシステムを適切に運用するためには、作業にコミットできる情報システムの高度な知識を持った者が 1 人は必要になると考える。また、今回行ったデータ圧縮等、HIC 利用におけるノウハウについても利用者に対する一定の情報提供が必要と考える。

HIC は AWS 上に構築されており、利用状況に応じてクラウド利用料が発生するシステムである。このため、利用者に対してはどの処理に対して費用がかかるかなどの説明がある方が良いと考える。この知識がないと利用方法によっては多大な費用が生じてしまい、HIC 運用が立ち行かなくなる可能性が懸念される。利用者リテラシー向上の観点からも費用に関する情報提供も重要と考える。

Ｅ．結論

HIC における NDB 特別抽出データ分析環境構築の課題を抽出し、対応策の検討を実施した。HIC においてある程度大きなデータを分析する場合はストレージ容量を考慮した対応等が必要となり、情報システムの高度な知識を持った利用者の参加が必要になると考える。

Ｆ．健康危険情報

なし

Ｇ．研究発表

なし

Ｈ．知的財産権の出願・登録状況

なし

図 1　HIC 利活用に関する問題（Windows 環境）