厚生労働行政推進調査事業費補助金(厚生労働科学特別研究事業)

(総括・分担) 研究報告書

Ⅱ．分担研究報告

1.クラウド解析基盤 HIC を活用した NDB 通年パネルデータセットの迅速提供プロセスと支援体制の実証評価

研究代表者　村松　圭司・産業医科大学　医学部・准教授

分担研究者　明神大也・国立大学法人浜松医科大　・健康社会医学講座・准教授
分担研究者　牧戸香詠子・東京大学大学院医学系研究科生物統計情報学講座・特任助教
分担研究者　森由希子・京都大学医学部附属病院　医療情報企画部・准教授

研究要旨：公的医療保険レセプト情報等を収載する NDB 通年パネルデータセットを対象に、申出・承諾・データ受領・データベース化・網羅的記述統計・成果公表までの全ライフサイクルをクラウド解析基盤 HIC 上で実行し、支援体制と技術運用上のボトルネックを検証した。2024 年 7 月 5 日の事前相談から 11 月 25 日の利用開始まで 143 日、承諾後 45 日で HIC に接続でき、NK（約 130 GB）は 8 営業日、ZK（約 220 GB）は 3 営業日で DB 化した。全テーブル・全カラムについて最小値・最大値・平均・標準偏差・欠損率、頻度上位 50 値を自動抽出し、汎用データディクショナリと PowerShell テンプレートを整備することで初学者の操作負荷を大幅に低減できることを確認した。結果、従来の特別抽出平均と比べ半期以上短いデータ提供が可能であり、仮想環境性能も十分であった。ログイン手順、マニュアル配置、診療報酬マスタ整備、FAQ 追補等の具体的改善策を提言し、迅速提供体制の実装と NDB 利活用拡大に資する実践的指針を示した。

A. 研究目的

匿名医療保険等関連情報データベース（NDB）は、公的医療保険レセプト情報を網羅的に収載する国内最大規模の医療ビッグデータであり、国民皆保険制度下における医療政策立案および臨床疫学研究に不可欠な基盤である。厚生労働省は令和 6 年秋を目途に、クラウド上で解析可能な医療・介護データ等解析基盤（HIC）を活用し、探索的解析に十分な症例数と 1 ～数年の縦断追跡を可能とする「通年パネルデータセット」を提供する新体制を構築しつつある。本研究は、この通年パネルデータセットを実際に申出・取得・解

析・成果物公表に至るまで一連のライフサイクルで運用し、研究者視点と提供者視点の双方から支援体制の実効性を検証することを目的とする。具体的には、①申出手続の可搬性とクラウド解析環境の操作性を評価し、②変数特性の網羅的把握とクエリ作成テンプレートを整備し、③得られた研究成果を通じて支援業務マニュアルおよび配置型教材に反映し、④迅速提供を阻害する技術的・運用的ボトルネックを抽出して提言する。これにより、令和 6 年秋に閣議決定された迅速提供体制の達成に寄与し、NDB 利活用促進のための実践的エビデンスを提示するものである。

B. 研究方法

B-1　研究デザインと全体フロー

本研究は、通年パネルデータセットを用いた横断研究である。今後の HIC 利活用促進に向けた資料を得るため、全テーブル・全カラムの集計を行う。はじめに、通年パネルデータセットを HIC 解析環境で利用する申請を行い、承諾が得られた後、各拠点で利用環境をセットアップする。続いて、提供されたデータをデータベース化し、集計を行う。最後に、最終生成物として集計表を、副生成物としてクエリを取り出す手続きを行う。なお、公表前確認は HIC 取り出し時に同時に行われる。

B-2　データセットの範囲と対象

2024 年度に利用可能であった通年パネルデータセット全データを対象とした。格納されている期間は 2022 年度の医科・DPC・歯科・調剤レセプト及び特定健診・特定保健指導である。テーブルやカラムの一覧は厚生労働省の Web サイト『二次利用ポータル』の「コンテンツ」から「利用を検討している方々へのマニュアル」内にある「環境別　利用者マニュアル」の「利用者マニュアル（トライアルデータセット・通年パネルデータセット）.zip」をダウンロードすることで閲覧可能である。

B-3　申出手続・データ取得プロトコル

通年パネルデータセットを利用するためには、『二次利用ポータル』の「各種申請」から、「新規利用申請」のうち、「利用申請(探索的利用環境)」を選択する。その後、遷移した画面で必要事項を入力すれば申請が完了する。なお、本研究では、二次利用ポータルの本格運用開始前に試行的申請として行ったため、事務局へのメールでの申請を行っている。また、本研究では Windows 環境を選択した。

B-4　HIC 解析基盤の構築

申請が受理された後、接続するための情報等が事務局から提供されるので、それに従って解析環境のセットアップを行った。HIC への接続後、既に配置されている通年パネルデータセットのテキストファイルをデータベースソフト(PostgreSQL)に取り込み、データベースを構築した。

B-5　探索的データ解析手法

本研究では、①各テーブルに対し、それぞれのカラムの最小値、最大値、平均、

標準偏差、欠損数、欠損割合、②もう一つは、各カラムに格納されている値の頻度が高いものから上位 50 の出現数の二種類の集計を行った。

B-6　クエリ作成・検証方法

集計に用いたクエリを別紙 1 に示す。HIC には PostgreSQL 用の GUI ツールがインストールされていないため、Windows に標準で搭載されている Powershell でデータベース操作を行った。成果物は、①テーブル単位、②カラム単位でテキストファイルとして書き出した。集計結果を別紙 2 に示す。この集計は、通年パネルデータセットを用いた研究を計画する際の、データディクショナリとしても機能する。

B-7　情報セキュリティと倫理的配慮

HIC の利用条件に従い、職員以外が立ち入らない部屋で操作を行った。また、探索的利用環境は事前の倫理審査が不要であるため、申請を行わなかった。

B-8　研究支援フィードバック収集

すべての作業を完了した後、データベース操作を担当した研究者を中心に、困難さが伴う作業や、初学者がつまずきやすい点について整理した。

C. 研究結果

C-1　データ取得および処理実績

事前相談を 2024 年 7 月 5 日に開始した。最終の書類調整は 8 月 19 日に完了した。9 月 4 日の匿名医療情報等の提供に関する専門委員会にて個別審査が行われ、10 月 11 日に厚生労働省から承諾された。その後、誓約書等をとりまとめ、10 月 31 日に依頼書を提出し、11 月 25 日から HIC の利用を開始した。事前相談から利用開始までは 143 日、承諾から利用開始までは 45 日であった。利用開始後、C-2 の処理開始するまでに、(1)HIC 解析環境へのログイン　(2)再 DB 化の実施　の 2 ステップを実施し、NK ファイルでは 9 営業日、NK ファイルで慣れたため ZK ファイルでは約 3 営業日を要した。概ね、厚生労働省及びその事業者が作成されたマニュアル等に従って比較的円滑に遂行できたと感じるが、その中で認識した課題と要望事項を列挙する。

(1) HIC 解析環境へのログイン

厚生労働省より、第三者提供窓口を経由して電子証明書と解析環境利用に関する通知書 PDF が送付される。その情報を用いて、HIC 解析環境操作端末にインストールした AWS コンソールアプリから Workspaces にログイン、その後 EC2 環境にリモートデスクトップ接続で接続する形式をとっている。

- 当初、HIC ポータルは手続担当者しかアカウント登録しておらず、他の取扱者は HIC ポータルにはログインできない状態であった。しかしながら解析環境利用者マニュアルは HIC ポータル上か Workspaces 内にしか存在せず、さらに HIC ポー

タル内の解析環境利用者マニュアルは抜粋版であった。その結果、迅速に利用者マニュアルを参照できず、HIC 解析環境へのログインに時間を要した。利用者マニュアルの配置方法を検討していただきたい。さらに「どのマニュアルから読めばよいのかよくわからない、まとめてほしいが Windows と Linux は別々に作成してほしいかも」という意見もあった。また、まったく本質的ではないが、利用者マニュアルにある google authenticator のアイコンが古かった。

- 解析環境利用に関する通知書の上部に EC2 名やログイン情報(OS ローカル管理者ユーザ)が記載されていたが、本研究では用いた経験がなかった。不要であれば除外するか、タイミングに応じて必要であれば、必要なタイミングを記載するとともに通知書下部に移動していただきたい。
- 解析環境利用に関する通知書 PDF でパスワードを Amazon WorkSpaces クライアントアプリケーション（以下、AWS クライアント）にコピー＆ペーストすると、毎回半角空白が混入してしまった。通知書 PDF の形式を再検討いただきたい。
- そのためにアカウントロックされてしまい、解除手続きが必要になった。さらに AWS クライアントでは、アカウントロックされても ID/PW の間違いでも、「ID/PW が無効です」という旨のメッセージが表示され、ID/PW 間違いなのかアカウントロックなのかが不明瞭であった。本事象を改善するか、AWS の仕様であればマニュアルまたは QA 等に明記いただきたい。
- Workspaces に入るとマウスポインタが表示されなくなる事象が発生した。窓口に問い合わせたところ、Workspaces 内のセキュリティアップデートに起因する事象との回答があり、Workspaces の restart 処理で解決したが、こういった事象に対する QA 構築が重要であると感じた。
- 解析環境利用に関する通知書に記載されている PW は原則、「初期パスワードは必ず変更いただきますようお願いいたします。」と記載されており、変更は当然だと思うが、Postgre へのログインパスワードのみ、HIC 解析環境利用者間で共用となっており、誰かが変えると利用者間で共有しないといけない。この点は何か要望を求めるものではないが、HIC 解析環境利用者が注意しなければならない点だと感じた。
- HIC 解析環境へのログイン時には二要素認証が必須であるにもかかわらず、パスワードの定期変更と直近 2 回のパスワードは使えないことになっている。パスワードルールが厳しいと感じるので、緩和をご検討いただきたい。
- HIC 解析環境へのログイン時には二要素認証が必須であるにもかかわらず、HIC ガイドラインで、HIC 解析環境操作端末へのログイン時にも二要素認証が求められている。HIC 解析環境操作端末ログイン時の二要素認証の必要性が乏しいと感じるので、緩和をご検討いただきたい。
- Workspaces 上に配置いただいている「別紙_01_コマンド一覧.xlsx」に「データ管

理番号」を入力するセルがあり、「データ管理番号」は通知書参照となっているが、通知書に「データ管理番号」の項目が存在しないので、改善をお願いしたい。

- 別紙 4 にテーブル定義書として各ファイルの情報が掲載されているが、シート名が長くよみづらかった。ユーザビリティの観点から改善を検討していただきたい。

(2) 再 DB 化の実施

HIC 解析環境の EC2 環境にログイン後、S3 に配置されている通年パネルデータセットの CSV ファイルから、PostgreSQL でデータベース（DB）化を行った。NK と ZK と存在し、NK が 130GB 程度、ZK が 220GB 程度であったが、NK の DB 化（copy クエリ実行完了まで）に 8 時間程度要した。

- 利用者マニュアルに DB 化手順が掲載されており、また Create 文や Copy 文のテンプレートを配置してくれているため、DB 化の手間はかなり省略できている。しかし根本的な課題提起になってしまうが、通年パネルデータセットは定型化されているため、研究者等への提供時点で PostgreSQL に DB 構築しておいてほしい。
- Workspaces 内に配置されている利用者マニュアルは Word で、実際に通年パネルデータセットを操作する EC2 環境には Word その他 Microsoft Office 関連がインストールされていなかったため、利用者マニュアルを読むためには Workspaces に戻る必要があった。利用者マニュアルは PDF のほうが読みやすいので、EC2 上から PDF で読めるように対応をお願いしたい。さらに、Microsoft Office がないと最終生成物または成果物の作成に手間がかかるため、EC2 環境でも Microsoft Office を利用できるようにしていただきたい。
- リモートデスクトップ接続のため、デフォルトではドライブはリモート接続されない。リモートデスクトップ接続に長けたユーザーであれば設定変更可能であるが、利用者マニュアルに EC2 環境でも Workspaces のドライブ参照できる設定変更を記載していただきたい。
- マスタファイルに特定健診実施機関及び特定保健指導実施機関のマスタが入っているが、そもそも thoken_entry ファイルの健診実施機関コードが削除されているため、当該マスタファイルは機能しないため削除を検討いただきたい。
- S3 からダウンロードした際に、TPDS_NK_REPORT.csv というファイルがあったが、別紙 4 のテーブル定義書には記載されていなかったため、テーブル定義書への追記または当該ファイルの削除を検討いただきたい。

その他困った点として、「1 月 27 日～31 日に WorkSpaces 及び EC2 のメンテナンスを行うので使わないでください」という旨の通知が 1 月 22 日に届いた。おそらく緊急メンテナンスだと思われるが、年度末の分析の佳境時に実質 1 週間 HIC 解析環境を使えなくなっ

たのは困った。どうしてもシステム稼働に必要なものだと思われるが、今後通年パネルデータセットをはじめとする HIC の利用者が増加した際は、ユーザー影響にも配慮いただきたい。

C-2　変数別記述統計サマリ

全テーブル・全カラムに対し、最小値・最大値・平均・標準偏差・欠損数・欠損割合を算出した。結果が膨大であるため、本報告書では以降、全数起点の医科診療行為の RE レコードを例に記載する。レセプト種別や診療年月、男女区分列は欠損率 0.0 % であった。郵便番号列では欠損率 2.1 % が認められた。頻度上位 50 値の抽出では、患者属性テーブルの男女区分列において、女性が 27,378,243 件、男性が 20,085,726 件であった。

C-3　支援フィードバックの整理

データベース操作を担当した研究者 2 名で課題を整理した。主な困難として「PowerShell 操作への不慣れ」「マスタ類の整備不足」の 2 点が挙げられた。また、初学者向けの改善点として、「ER 図を整備すること」、「パターン別に統計解析用に準備するテーブルレイアウトを示すこと」の 2 点が挙げられた。

D. 考察

探索的利用環境であったこと、HIC 環境での利用であることで、申請から利用開始までの期間は NDB 特別抽出の平均期間と比較し半分以下となった。また、全テーブル、全カラムを集計する負荷の高い作業を行ったが、特段の問題なく完了したことから、準備される仮想環境の性能も十分であると考えられた。ただし、本研究では SQL 上で集計を行い、テキストに書き出しているため、R 等の統計ソフトでの挙動は検証していない点に留意が必要である。今後の利用者の支援について、「PowerShell 操作への不慣れ」「マスタ類の整備不足」の 2 点が挙げられたが、このうち前者は FAQ の追補およびサンプルスクリプトの追記で対応可能であり、後者は診療報酬マスタの併置により解決可能であると考えられた。マスタは『二次利用ポータル』内の「コンテンツ」に「マスタ共有」機能が存在する。今後の研究者間の互助が期待される。また、初学者向け推奨手順として、「ER 図を整備すること」、「パターン別に統計解析用に準備するテーブルレイアウトを示すこと」の 2 点が挙げられた。前者については、NDB を利用するためにはレセプトデータの構造そのものへの理解が必要であることが制度開始当初から繰り返し言及されているが、広く用いられている成書は本研究の研究者の知る限り存在しない。全国共通で利用できる教科書のようなドキュメントは NDB 利活用促進に向けて根幹となる資料と考え、今後の整備が期待される。後者については、既に利用を終了した申請の成果物を整理することで、NDB を用いて実施できる研究の類型化が可能であると考えられた。

本研究は複数拠点で実施したが、各拠点の作業状況を随時共有する必要があったもの

の、各拠点の研究者の任意の時間帯に。これまでは 1 拠点で DB 構築から分析、集計結果作成まで実施する必要があったが、複数拠点で同じデータを扱うことができる点で、クラウド環境である HIC を使う最大のメリットを感じられた。特に DB 操作者が複数拠点にいたり、DB 構築者と NDB 操作者（DB 後のデータを分析する者）が別拠点にいたりする場合は、HIC を使うことが便利と感じられた。

E. 結論

本研究は、通年パネルデータセットを HIC 上で一連のライフサイクルで運用し、申請から利用開始までを平均半期以下に短縮できること、仮想環境性能が全テーブル横断集計に耐えることを実証した。また、操作習熟とマスタ整備を進めれば初学者の障壁は大幅に低減できる見通しを得た。これらの知見は、迅速提供体制の確立と NDB 利活用拡大に資する実践的指針となる。

F．健康危険情報

なし

G．研究発表

1. 論文発表

なし

2. 学会発表

なし

H．知的財産権の出願・登録状況

（予定を含む。）

1. 特許取得

なし

2. 実用新案登録

なし

3.その他

なし