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、初学者がつまづ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. 現状の課題

1.　データ提供・手続面

1 申請フローの分散

試行期にはメール等の個別連絡が併存し、二次利用ポータルへの集約が不十分である。

2 タイムラインの可視性不足

申請～許可～利用開始までの標準所要日数・ボトルネックの情報公開が限定的である。

3 環境差異の事前明示不足

探索的利用環境と本申請環境でのテーブル構成・件数差が計画段階で把握しにくい。

2 解析基盤・運用面（HIC）

1 GUI 不在に伴う操作習熟負荷

PowerShell 等のテキスト操作依存が初学者に高い学習コストを強いる。

2 自動化・再現性支援の不足

標準スクリプト雛形やジョブ実行テンプレートが体系化されていない。

3 運用メトリクスの不十分さ

典型的クエリの実行時間や失敗率など性能指標が利用者から参照しにくい。

3 データ記述・マスタ整備

1 データ辞書の機械可読性不足

PDF 主体の提供であり、クエリ自動生成や検証の基盤として活用しにくい。

2 診療報酬・施設基準等マスタの対照性不足

コード対訳や履歴管理が断片的で、表記ゆれ対応が困難である。

3 横断的 ER 図の未整備

テーブル間リレーションの俯瞰が難しく、JOIN 設計に時間を要する。

4 教育資材・支援体制

1 学習資材の段階設計の不足

FAQ は整備途上であり、e-lerning の標準カリキュラムが未確立である。

2 支援窓口の役割定義の曖昧さ

システム照会と研究設計相談の切り分けが不十分で、回答者の専門性ミスマッチが発生する。

3 成果物の知識化不足

既存申請の成果物が類型化・再利用可能な知識として蓄積・公開されていない。

5 倫理・ガバナンス

1 倫理手続の適用範囲の周知不足

探索的利用における倫理審査要否の判断基準が利用者間で不

統一である。

2 ログ・監査情報の利用者提示不足

セキュリティ遵守の可視化（アクセスログの自己点検等）の仕組みが限定的である。

2 改善すべき点（提言）

1 データ提供・手続の一元化・透明化

1 申請フローの完全集約

二次利用ポータルへの手続統合と、メール運用の廃止計画を明示する。

2 標準タイムラインの公開

各段階の目標所要日数と想定遅延要因をダッシュボードで提示する。

3 環境差異の仕様書化

探索・本申請の差分表（テーブル・件数・利用制約）を事前公開する。

2 解析基盤の利用容易性と再現性の強化

1 GUI オプションの提供

軽量 GUI を選択肢として提供し、初学者の参入障壁を低減する。

2 テンプレート群の標準化

記述統計・頻度上位抽出・品質チェックの SQL／PowerShell 雛形を公式配布する。

3 性能メトリクスの可視化

代表クエリの実行時間分布・失敗率等を匿名化統計で定期公開する。

3 データ記述資源の機械可読化・体系化

1 データ辞書の多形式提供

PDF に加え JSON／CSV を提供し、PI 参照も可能にする。

2 マスタの統合管理

診療報酬・施設基準等のコード対訳・履歴をリポジトリ化し、表記ゆれ辞書を配布する。

3 公式 ER 図の公開

テーブル関係の標準 ER 図と変更履歴をバージョン管理で提供する。

4 教育資材と支援窓口の機能分化・連携

1 段階的カリキュラムの構築

初級（環境操作）—中級（JOIN 設計）—上級（研究設計）の e-lerning を整備する。

2 窓口の役割定義

システム照会は運用窓口、研究設計相談は専門家窓口とし、SL と KPI を別建てで設定する。

3 成果物の知識ベース化

申請終了案件のクエリ雛形・テーブルレイアウト・落とし穴をメタデータ付で再利用可能化する。

5 倫理・ガバナンスの実装強化

1 倫理手続の判定フロー提示

利用目的別の倫理審査要否判定チャートと事例集を公開する。

2 監査可能性の向上

利用者が自ら確認できるアクセスログ閲覧機能と自己点検チェックリストを提供する。

3 海外事例に基づく標準整備

教育資材・支援・品質管理の国際的ベストプラクティスを踏まえ、国内規範を段階導入する。

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