

I. 総括研究報告

NDB 研究支援体制の実践的検証研究

研究代表者 村松 圭司 産業医科大学 医学部 准教授

研究要旨：令和6年秋に本格提供が始まったNDB通年パネルデータセットをクラウド解析基盤HIC上で一気通貫に運用し、迅速提供体制の実効性と課題を検証した。事前相談から143日、承諾後45日で利用を開始し、総約350GBのCSVをPowerShell スクリプトで取込、NKは約9営業日、ZKは約3営業日でデータベース化した。全テーブルの記述統計を自動生成するテンプレートを整備し、初学者の解析障壁を低減した。一方、申請フォームの一時保存不可や質問ルーティングの不透明さなどHICポータル運用課題を確認し、下書き保存機能追加など具体的改善策を提言した。台湾 NHIRD・韓国 NHIS・英国 CPRD との比較から、層別抽出サンプルやライセンス制など多様なアクセス形態が国際共同研究を促進することを示した。迅速提供体制の最適化と支援資材の共有は、政策立案と臨床疫学研究の加速に寄与すると考えられた。

研究分担者氏名・所属機関名称・所属機関における職名

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

明神大也・国立大学法人浜松医科大学・健康社会医学講座・准教授

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

A. 研究目的

本研究は、令和6年秋より本格運用が始まったNDB「通年パネルデータセット」をクラウド解析基盤HIC上で申請・受領・データベース化・解析・成果公表まで一気通貫で実施し、その迅速提供体制と研究支援サービスの実効性を研究者・提供者双方の視点から総合的に検証することを目的とする。具体的には、①オンライン申請手続とHIC仮想環境の操作性を評価し、従来より短い手続期間でのデータ利用開始が可能かを明らかにする、②全テーブル・全カラムの記述統計を自動生成してデータディクショナリとPowerShellテンプレートを整備し、初学者が躓きやすい技術的ボトルネックを抽出・対策する、③HICポータルの掲示板・メッセージ機能を用いた実地運用を通じ、質問割り振りやマニュアル配置など運用上の課題を可視化し具体的改善策を提案する、④台湾NH

I R D・韓国NH I S・英国C P R D等の長期追跡サンプルデータセットの提供実態を比較調査し、日本版サンプル拡充と国際共同研究プラットフォーム構築の指針を提示する。これら4点を通じて、N D B利活用の障壁を低減し、医療政策立案および臨床疫学研究の加速に資する実践的エビデンスを創出することを目指す。

B. 研究方法

本研究は ①国内 NDB 通年パネルデータセットの実地運用評価、②HIC ポータルの運用課題解析、③国際的サンプルデータ提供事例の比較調査——の三相から成る混合法研究として設計した。

1. NDB 通年パネルデータセット実地運用

2022 年度分の医科・DPC・歯科・調剤レセプト、特定健診・特定保健指導を含む通年パネル全データを対象とし、二次利用ポータル「利用申請（探索的利用環境）」経由で申出を行った。試行段階につきメール申請を併用し、事前相談開始（7 月 5 日）から利用開始（11 月 25 日）まで 143 日、審査承諾から 45 日で接続を完了した。接続後、HIC EC2 環境に配置された CSV（NK 約 130 GB、ZK 約 220 GB）を PostgreSQL に取り込み、NK は COPY 実行完了まで 9 営業日、ZK は 3 営業日で DB 化を終えた。GUI がないため DB 操作は PowerShell で行い、CREATE/COPY 文をテンプレート化して作業時間を短縮した。

2. 探索的データ解析手順

全テーブル・全カラムについて（1）最小値・最大値・平均・標準偏差・欠損数・欠損割合、（2）出現頻度上位 50 値——を自動集計し、テキストとして書き出した。生成物はデータディクショナリ兼クエリ雛形として再利用可能である。加えて ER 図作成とテーブルレイアウト例を提示し、初学者向け推奨手順を整理した。操作は HIC ガイドラインに基づき施錠室内で実施し、探索的環境のため倫理審査は不要と判断した。

3. HIC ポータル運用課題の解析

ユーザー登録、オンライン申請、掲示板・メッセージ投稿を令和 6 年 11 月に実施し、操作性、一時保存機能、複数施設入力制限、通知フローなどの課題を抽出した。質問機能は掲示板＝全体共有、メッセージ＝個別対応と仕様付けられているが、区分が不明瞭で利用者の混乱が確認された。各投稿に対する回答速度と回答主体（支払基金・厚労省・窓口事業者等）の適切性を記録し、改善提案を整理した。

4. 海外サンプルデータ提供事例の比較調査

HIC での追跡型サンプル拡充の示唆を得るため、文献レビューと関係機関ヒアリング

を実施した。台湾 NHIRD（約 200 万人、施設内オンサイト、IRB 不要）、韓国 NHIS サンプル（約 100 万人、OPEN COHORT、リモート可、6 か月使用）、英国 CPRD（人口の 0.02-0.13%層別抽出、長期縦断、年額ライセンス制）を対象とし、抽出設計、追跡期間、アクセス条件、倫理要件を比較した。なお、フィンランドにおいては、COVID-19 患者をはじめとするの長期コホートデータセットの提供をリモートアクセスにて行っており、通常のデータ抽出と比較して迅速な提供と低コストを実現させているが、サンプルデータは存在しなかった。

5. 研究支援フィードバック

全工程終了後、複数拠点の DB 操作者と解析担当者から作業負荷・操作難易度を聞き取り、課題を「PowerShell 操作への不慣れ」「診療報酬マスタ整備不足」に集約した。前者は FAQ 追補とサンプルスクリプトで、後者はマスタ共有機能の活用で解決可能と評価した。

以上により、現行 HIC 迅速提供プロセスの所要時間、仮想環境性能、支援資材、ポータル UI／ガバナンス、および海外事例との比較指標を包括的に取得した。

C. 研究結果

1. 通年パネルデータセット取得・処理実績

事前相談開始（2024 年 7 月 5 日）から利用開始（11 月 25 日）まで 143 日、厚生労働省の承諾後 45 日で HIC 解析環境への接続が完了した。HIC EC2 上に配置された CSV は NK 約 130 GB・ZK 約 220 GB で、PowerShell テンプレートによる COPY 実行により NK は 9 営業日、ZK は操作習熟後 3 営業日で DB 化を終えた。ログイン～DB 構築までの 2 工程に要した作業は、NK で 9 営業日、ZK で 3 営業日と記録され、テンプレートとマニュアル配置が処理時間短縮に寄与した。

2. 網羅的記述統計・テンプレート整備

全テーブル・全カラムについて①最小値・最大値・平均・標準偏差・欠損数・欠損割合、②出現頻度上位 50 値を自動抽出し、テキスト出力した。これによりデータディクショナリと再利用可能な SQL 雛形を生成し、初学者の変数探索を大幅に効率化できることを確認した。また ER 図と代表的テーブルレイアウトを作成し、解析パターン別の推奨手順を整理した。クラウド環境で同一データを複数拠点が同時操作できたことは、従来のローカル環境にはないメリットであり、DB 構築担当と解析担当が離れた施設に在籍する場合の作業分担を容易にした。

3. HIC ポータル運用評価

オンライン申請では「運用管理規定等が1施設分しか入力できない」「申請フォームが一時保存できない」など構造的制約が判明した。掲示板とメッセージの使い分けが周知不足で、通知先が担当者限定される仕様も進捗共有を阻害した。さらにユーザー登録には厚労省側承諾を要するため数時間～数日を要し、申請ドラフト機能の欠如が書類決裁プロセスを煩雑にしていた。これらの課題を基に、①複数施設入力フィールドの追加、②下書き保存・PDF出力機能の実装、③質問割当ての自動分類アルゴリズム導入を提案した。

4. 解析環境の課題と改善余地

利用者は PowerShell 操作への不慣れと診療報酬マスタ欠落を主な障壁に挙げたが、前者は FAQ 追補とサンプルスクリプトで、後者は「マスタ共有」機能の活用で解決可能と評価された。また、マニュアルが Word 形式で WorkSpaces 側にのみ配置されていたため、EC2 環境から閲覧しにくいこと、緊急メンテナンスによる計算停止が年度末の解析に影響したこと等が報告された。

5. 海外追跡サンプルとの比較

台湾 NHIRD は約 200 万人（人口の 8.7–9.4%）を層別ランダム抽出したサンプルを 2000・2005・2010 年起点で提供し、IRB 不要でオンサイト解析が可能。

韓国 NHIS サンプル は全人口の 2%に相当する約 100 万人を層別抽出し、OPEN COHORT 方式で新生児を逐次追加、指定施設またはリモートで 6 か月間利用できる。

英国 CPRD は人口の 0.02–0.13%を対象とする Aurum/GOLD 両サンプルを提供し、長期縦断追跡と年額ライセンス制を特徴とする。

これらと比較すると、日本版通年パネルは全量データを基にした固定コホートである点で網羅性が高い一方、年度拡張時の追跡設計とアクセス形態の多様化（リモート利用・年更新ライセンス等）が課題として浮上した。

6. 総合的知見

- (1) 申請～利用開始期間を半期末満に短縮でき、仮想環境は全テーブル横断集計を問題なく処理できた。
- (2) 自動集計スクリプトとデータディクショナリは HIC 内共有資材として再利用価値が高く、初学者支援に直結する。
- (3) HIC ポータルは申請 UI・情報共有・質問割当ての 3 点で機能改善余地が大きい。
- (4) 海外 3 事例は層別抽出・ライセンス制・OPEN COHORT など多様な提供形態を採用しており、日本版サンプル開発や国際比較研究の設計指針を与える。

以上の結果は、迅速提供体制の実装と NDB 利活用拡大に向けた技術・運用両面の具体的改善点を明確に示した。

D. 考察

本研究は NDB 通年パネルの「迅速提供体制」を実地で検証し、技術的・運用的両面の課題と改善点を多層的に抽出した。まず、承諾後 45 日での利用開始は従来の特別抽出（平均 200～240 日）に比して半期以上の短縮であり、申請手続と仮想環境性能が十分に実効性を備えつつあることを示した。一方で CSV→DB 化に要する作業は PowerShell 操作と診療報酬マスタ整備に依存し、操作スクリプトやマスタ共有を常設化しなければデータ科学に不慣れな研究者の参入は限定的となる。今回整備したテンプレートとデータディクショナリを HIC 内の共通資産として公開すれば、解析初期のつまずきを最小化し、複数施設が並行して同一データを活用する協働研究も促進できる点は大きな利点である。

HIC ポータルの UI／ガバナンスについては、①複数施設入力不可、②申請フォームの一時保存不可、③掲示板・メッセージ区分の不透明さ——が運用負荷を高めていた。これらは入力フィールドの拡張と下書き保存機能、質問自動仕分けアルゴリズムの導入で改善可能と判断した。特に掲示板質問の回答主体が部署により分散している現状は回答速度と責任所在の両面で課題であり、タグ付けによるカテゴリ別ルーティングが有効と考えられる。

海外 3 事例との比較では、台湾 NHIRD が大規模固定コホートをオンサイト解析として供し、韓国 NHIS サンプルが OPEN COHORT 方式で追跡対象を逐次追加、英国 CPRD が年額ライセンス制と人口カバー率向上を両立させている。これらを踏まえると、日本版通年パネルの将来像として、①層別抽出サンプルを追加しオンデマンド配布を可能にする、②OPEN COHORT に近い年度更新と死亡・転出者補完を実装する、③年間ライセンス制＋リモート利用を検討し国際比較の参入障壁を下げる——など多様なアクセス形態が求められる。また、大規模データセットの迅速提供は学術利用に留まらず、自治体の医療計画策定や保険者の保健事業評価など政策実装の即応性向上にも寄与し得る。

最後に限界点として、今回の実地検証は単年度・単一研究グループでのプロセス測定であり、施設種別・研究目的の多様性を十分に代表していない。今後は複数グループが同一テンプレートを用いて再現試験を行い、処理時間・障壁要因の一般化可能性を検証する必要がある。それでも本研究で得た実測値と改善提案は、HIC の運用高度化ロードマップの基礎資料として有用であると考えられる。

E. 結論

NDB 通年パネルは承諾後 45 日で利用開始でき、仮想環境は全表横断解析に耐えうる性能を有した。テンプレートとデータディクショナリの共有により初学者の参入障壁は大

幅に低減する。一方、HIC ポータル UI と質問ルーティングには改善余地が残り、海外事例に学ぶ柔軟なアクセス形態の導入が今後の課題と考えられた。

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