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））
総 括 研 究 報 告 書（令和 7 年度）

# NDB のユーザビリティ向上を通じて クラウド上でのデータ二次利用を推進するための研究

研究代表者 明神 大也（浜松医科大学　健康社会医学講座）

**研究要旨**

**背景と目的:** NDB は、レセプト情報や特定健診等情報を格納した公的データベースであり、第三者提供の法制化、死亡情報等の収載拡大、HIC の運用開始により、利活用の基盤整備が進められている。一方で、プリセットデータが CSV 形式で提供されることや、解析環境でコマンドライン操作が必要であることなど、ユーザビリティには課題が残る。本研究では、ユースケース調査、プリセットデータ改善、クラウド環境の操作性向上、技術供与等を通じて、NDB/HIC の利便性を高め、医療等情報の二次利用を推進することを目的とした

**方法:** NDB-β、通年パネルデータセット等の利用申請を行い、HIC 上での活用可能性を検証した。具体的には、過去の NDB 提供申出成果物を対象としたユースケース調査、HIC/NDB ガイドラインおよび利用者マニュアルに基づくユーザビリティテスト、標準化された SQL データハンドリング手法の整理、S3・Parquet・DuckDB を用いたクラウド分析基盤の検証を行った。さらに、韓国・フィンランド・スウェーデン・フランスの医療関連データ提供状況を調査し、研究用マスタ情報の共有を試みた。

**結果:** 論文 47 件、報告書 16 件を分析した結果、論文 24 件、報告書 6 件は HIC で提供可能なデータ形式により実施可能と考えられた。一方、希少疾患や詳細な層別解析では特別抽出が必要であった。HIC の解析環境やポータル等には改善課題が確認された。SQL 処理では、解析単位、時間的アンカー、観察ウィンドウ等に基づく標準ワークフローを整理した。また、1 年分 620GB のデータを S3 上に Parquet 化し、低コストで実用的な分析基盤の可能性を確認した。

**考察:** 本研究により、NDB/HIC の利活用拡大には、汎用データセットの拡充、申請様式やポータルの操作性改善、再利用可能な SQL 処理手法、低コストで柔軟なクラウド分析基盤、マスタ共有の促進が重要であることが示された。特に通年パネルデータセットの複数年化は、既存研究の一定割合を HIC 上で代替可能にする可能性がある。また、S3・Parquet・DuckDB を組み合わせた基盤は、従来の Redshift 依存によるコストや柔軟性の課題を補完し得る。

**研究分担者**

村松 圭司（千葉大学医学部附属病院）
森 由希子（京都大学医学部附属病院）
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）
松居 宏樹 （東京大学大学院医学系研究科）
杉山 雄大（国立健康危機管理研究機構・国立国際医療研究所糖尿病情報センター）
柏木 公一 （国立健康危機管理研究機構・国立看護大学校）
西岡 祐一（奈良県立医科大学）

**研究協力者**

川田 加奈子（浜松医科大学　健康社会医学講座）
古野 孝志（国立健康危機管理研究機構・国際医療協力局グローバルヘルス政策研究センター）
市瀬 雄一（国立健康危機管理研究機構・国際医療協力局グローバルヘルス政策研究センター）
加藤 源太（京都大学医学部附属病院　診療報酬センター／病床運営管理部）

## A．背景と目的

匿名医療保険等関連情報データベース（以下「NDB」という。）は、厚生労働省が法令に基づき収集・提供しているレセプト情報や特定健診・特定保健指導情報などを格納したデータベースである。2020 年より民間事業者を含む第三者への提供が法制化されて以降、他の医療・介護データ等との連結解析、死亡情報等収載情報の拡大が進んでいる。また、医療ＤＸの推進に関する工程表（令和５年６月２日医療 DX 推進本部決定）、規制改革実施計画（令和５年６月 16 日閣議決定）「NDB や公的統計データの利活用の円滑化・迅速化」に従って、オンライン上で利用申請や審査が可能となる二次利用ポータルおよびクラウド上の解析基盤として「医療・介護データ等解析基盤（HIC）」の運用も開始され、データの迅速提供・利活用促進のための提供体制整備が進められてきた。HIC で迅速提供できるデータ形式は、サンプリングされた一年間のデータである「通年パネルデータセット」、サンプリングされた単月毎の断続データである「トライアルデータセット」、機微コード等をマスクしたレセプト全件を含む「NDB-β」の３種類とされている（2026 年５月時点）。しかしプリセットデータがクラウド上にも関わらず CSV ファイルで提供される、解析環境内ではコマンドラインの操作が必要である等、これまでのオンプレミス環境における解析環境と比較しても、ユーザビリティに関して多くの課題が残っている。

こうした中で、規制改革推進に関する中間答申（令和６年 12 月 25 日）においては、公的 DB の仮名化情報について安全であるのみならず迅速かつ円滑に利用・解析を行うことができるクラウド環境について、HIC との関係を整理しつつ実現することとされている。

そこで本研究では、ユースケース調査、プリセットデータの改善提案、クラウド環境の操作性への提言、技術供与等を通じて、NDB/HIC のユーザビリティを向上し、医療等情報の二次利用を推進することを目的とする。

具体的には①ユースケース調査、②現行の HIC のユーザビリティテストと改善提案、③汎用性の高いマスタや SQL クエリ等のプログラム作成、④汎用性の高いデータセットの検討・具体的な仕様作成、⑤利便性の高いクラウド環境の検討・提案、⑥HIC を利用して解析可能なリサーチクエスチョンやその手法の例示、⑦研究者支援体制への協力それぞれの分野で研究を進める。

## B．研究方法

本研究は、上記を満たすために、以下の A～F の研究を実施した。その前提として、NDB-βのオンプレミス環境での提供申出と、通年パネルデータセット及び NDB-β の HIC 上での利用申請を行った。

A) ユースケース調査と汎用データセットの検討

HIC で利用可能なデータセットの幅を広げるために、ユースケース調査をおこなった。具体的には NDB データの第三者提供に係る二次利用ポータルで公開されている「過去の提供申出の成果物」から、発表形式が論文または報告書（厚労科研に限る）に分類されている既存研究を対象とした。対象研究について、利用目的、分析に必要なデータ項目、対象疾患等を整理し、HIC で提供可能なデータ形式に基づくユースケースの分類を行った。さらに、分類結果を踏まえ、HIC 利用申請時に必要となる別添８「申出依頼テンプレート（抽出）」の記載例を作成し、具体的なデータ仕様の検討を行った。これは背景にある①と④に該当する。

B) HIC のユーザビリティテストと改善提案

「匿名医療保険等関連情報データベース（NDB）の利用に関するガイドライン」及び「医療・介護データ等解析基盤（HIC）の利

用に関するガイドライン」（以下「HIC/NDBガイドライン」という。）、及び二次利用ポータルに掲載されている利用者マニュアルの記載に従って HIC のユーザビリティを確認し、HIC 解析環境・二次利用ポータル・関連ドキュメントに改善余地のある課題を発見した。これは背景にある②に該当する。

C) 標準化された SQL データハンドリングフレームワークの提示

データの取り扱いには一般に SQL を用いたデータハンドリングが行われる。しかし、SQL は記述の自由度が高いゆえに、研究者ごとに独自のクエリが書かれる傾向がある。そのため、クエリ設計の方針が担当者の経験と裁量に依存し、コードが引き継がれる際に意図が伝わらず、研究の再現性が損なわれることも少なくない。そのため、SQL のクエリを共有しても、そこには膨大な量のレビュープロセスが発生し、第三者によるコードレビューが困難である。また、エラーが潜在しても発見が難しい。また、数テラバイトに及ぶ元テーブルを不必要に全件スキャンしたり、同一内容のテーブルを重複してコピーしたりするクエリが散見され、計算資源を浪費する原因ともなっている。

大規模時系列医療データを疫学研究に使用しやすい形に整理するための、標準化されたデータハンドリングプロセスを一般化可能なように提示することである。具体的には、(1) 解析単位、(2) 観察のタイムライン、(3) マスタ整備という三要素を軸としたクエリ設計の考え方を示し、各ステップを体系的に解説した。これは背景にある③に該当する。

D) 利便性の高いクラウド環境の検討・提案

NDB における HIC では、Redshift・ローカル Postgres 等の解析環境が提供されている。一方、同環境においては、Redshift 永続テーブル作成権限の欠如、ローカル解析サーバのメモリ・ストレージ不足、外部参照データ取込の困難、計算コストの増大といった複合的課題が存在する。

そこで HIC 内で提供されるクラウドストレージ(Amazon S3)と列指向ファイル形式 Parquet、軽量分析エンジン DuckDB を組み合わせた分析基盤を構築し、その実用性を検証した。これは背景にある⑤に該当する。

E) 諸外国の医療関連データ利活用・提供状況の調査

海外の医療データ提供の運用状況について調査（韓国・フィンランド・スウェーデン・フランス）を行った。その状況を踏まえ、解析環境の改善点、データ提供形式、マニュアル等の在り方等について検討し、利便性向上に資する提案を取りまとめる。これは背景にある①④⑤に該当する。

F) マスタ情報の提供

背景③にある汎用性の高いマスタが求められている事情として、NDB に含まれるレセプト情報は、医療機関が原則審査支払機関に提出したレセプトが、審査を経て、原則匿名化処理のみ行ったものであることが挙げられる。そのためいわゆる「コードの塊」になっており、研究テーマに沿った処理を研究者自身で行う必要がある。しかしながら研究テーマは研究者によって異なることから、多種多様なマスタが必要となるが、汎用性を示す根拠は乏しい。そこで二次利用ポータル上にある［コンテンツ］－［マスタ共有］を使いマスタ情報等の共有ができたり、研究者同士で誤りを指摘したりすることを目指すために、本研究担当者が NDB 等の研究で用いているマスタをアップロードすることとした。

## C．研究結果

上記方法に示した結果を下記に示す。なお各種 NDB の申出については、第 29 回匿名医療情報等の提供に関する専門委員会にて NDB-β のオンプレミス環境の提供承諾を、第 34 回匿名医療情報等の提供に関する専門委員会にて

通年パネルデータセット及びNDB-βのHIC上での利用承諾（いわゆる迅速提供）を得た。

A) ユースケース調査と汎用データセットの検討

本研究の分析対象は論文47件、報告書16件であった。論文では処方動向、介入効果（治療）、疾患疫学、地域差分析に関する研究が多く、報告書では地域差分析、医療提供体制調査、処方動向に関する研究が多かった。HICで提供可能なデータ形式に基づき分類した結果、論文24件（51.1%）、報告書6件（37.5%）が、通年パネルデータセット、トライアルデータセット、NDB-βのいずれかを用いた研究が実施可能であると考えられた。一方で、希少疾患や詳細な層別解析を伴う研究についてはサンプリングデータセットでは十分な分析対象数の確保が困難であり、特別抽出による提供が妥当と考えられた。（分担研究報告書１）

B) HICのユーザビリティテストと改善提案

HIC/NDB ガイドライン及び利用者マニュアルの記載に従って、HICにて通年パネルデータセット及びNDB-βを利用する際のユーザビリティの確認を行った。そして解析環境・二次利用ポータル・ドキュメントにかかる課題及び改善提案を一覧化した。(分担研究報告書２)

C) 標準化されたSQLデータハンドリングフレームワークの提示

「解析単位」「時間的アンカー」「観察ウィンドウ」「マスタテーブル整備」を基盤概念とし、研究デザインと SQL 構造を体系的に接続する方法を整理した。具体的には、解析単位テーブル作成、可変アンカー付加、時系列データ抽出、観察ウィンドウ集計、Long-to-Wide 変換という標準ワークフローを構築し、効率的かつ再利用可能なデータ処理を可能とした。この手法により、数TB規模の医療データに対しても計算効率を高めつつ、研究品質と教育的価値を向上できる。さらに、本フレームワークは複数の医療データベースへ横展開可能であり、将来的な研究基盤標準化や自動化にも資する汎用的手法である。（分担研究報告書３）

D) 利便性の高いクラウド環境の検討・提案

2024 年 4 月から 2025 年 3 月の医科・DPC・調剤の全テーブル(1年分、計620GB)をRedshiftからS3上にParquetとして生成し、MD5 ハッシュの加算突合により 684 テーブルすべてについてデータの一致を確認した。加えて、全テーブルの各カラムについて、NULL数、ユニーク値数、出現頻度上位50件、最大・最小・平均・標準偏差からなるデータサマリを生成した。HIC の甲区分(メモリ 64GB)の環境において、本構成が大規模レセプトデータに対する実用的な処理性能を有することを確認した。ストレージ料金も 1 年分 620GB で月額約 2,100 円と低コストで運用でき、従来の Redshift 依存の分析環境におけるコスト増大・柔軟性の不足といった課題を解消できる可能性を示した。(分担研究報告書４)

E) 諸外国の医療関連データ利活用・提供状況の調査

韓国では、NHID(Korean National Health Information Database)やHIRA(Health Insurance Review and Assessment Service)といった大規模データベースが整備され、オンサイトセンターを中心とした運用のもと、NHIDでは年間約1,000件のデータ提供の実績がある。フィンランドでは、Findataを通じた一元的な提供体制およびリモート解析環境が整備され、レディメイドデータセットの提供により迅速化とコスト削減が図られている。フランスでは、Health Data Hubを介して、レセプト等のデータベースである SNDS と希少疾患のレジストリである BNDMR の連結が進められている。スウェーデンでは個人番号制度を基盤として、行政・医療・福祉分野のレジストリが多層的に整備

されている。（分担研究報告書５）

F)　マスタ情報の提供

本研究の代表・分担研究者が NDB 研究等で用いており、他の研究者等の了解を得た薬価コードの時系列マスタ、新型コロナウイルス感染症（以下、Covid-19）における医療需要検討マスタ、がんにおける医療需要マスタ、糖尿病薬マスタ、病棟区分マスタの５マスタを提供した。前３つは GitHub 上に公開している。

薬価コードの時系列マスタは、社会保険診療報酬支払基金が公開しているマスタを再帰的にダウンロードし、日にちに応じた薬価、診療報酬を与える辞書型の構造として再構成している。Covid-19 における医療需要検討マスタは、日本で Covid-19 の医療資源利用状況を分析した研究で用いたものである。がんにおける医療需要マスタは、日本におけるがん治療のための医療資源利用状況を分析した研究で用いたものである。糖尿病薬マスタは 2026 年 3 月末時点のもので、日本でこれまでに販売された糖尿病関連治療薬の医薬品コード・商品名・一般名・成分・YJ コードである。病棟区分マスタはレセプトで算定できる入院基本料・管理料等の診療行為コードから、医療法上または診療報酬上どの区分に該当するかを示したものである（2013 年～2025 年）。

いずれも研究者が作成したものなので、信頼性を完全には保証できないが、一定のオーソライズは取られている。

## D．考察

本研究はユースケース調査、プリセットデータの改善提案、クラウド環境の操作性への提言、技術供与等を通じて、NDB/HIC のユーザビリティを向上し、医療等情報の二次利用を推進することを目的として実施された。その中で、本報告書では①～⑤を示した。

A)ユースケース調査と汎用データセットの検討については、ユースケース調査の結果から、現在１年間で提供されている通年パネルデータセットを５年以上に拡張することで、論文発表においては半数程度、厚労科研においても３分の１程度は HIC で提供可能なデータ形式で代用可能と考えられた。また、NDB の提供申出書様式１別添 8 は多数の項目が入っているため、研究者にとって操作が困難である。そこで、一定の条件下で妥当なテンプレート作成を行った。決してこれが全ての研究に適応できるわけではないが、今後の研究者等にとって参考になることを期待したい。また現在〇×で選択することになっているが、チェックボックスの方がユーザビリティは向上すると考えられる。

B)HIC のユーザビリティテストと改善提案については、本分担報告書に記載された内容を、今後の HIC の改善または情報連携基盤での仕様検討に反映されることを期待したい。

C)標準化された SQL データハンドリングフレームワークの提示に関しては、一定の限界があるものの、研究デザイン・デザインダイアグラム・SQL クエリ構造を一貫した論理構造として接続できることが強みである。将来的な SQL テンプレート自動生成、共通データモデルとの接続、AI 支援型クエリ生成、研究プロトコル自動実装などへの発展可能性を有していると考えられる。

D)利便性の高いクラウド環境の検討・提案に関しても、従来の Redshift 依存の分析環境におけるコスト増大や柔軟性の不足といった課題を解消し、低コストかつ再現性の高い大規模レセプトデータ分析基盤を実現できたと考えられる。セキュリティと利便性、コスト、瑕疵担保責任の４方向からバランスをとるのは困難であるが、より良い環境での分析に期待したい。

E)諸外国の医療関連データで活用提供状況の調査に関しては、今回、韓国・フィンランド・フランス・スウェーデンを調査した。特に、韓国とフィンランドに関しては、2024 年に調査を実施した際より、データ利活用環境が整備され

ていた。世界的にリアルワールドデータの整備の流れに沿って、レセプト関連データの利活用が進んでいることが実感した。フランスは、Health Data Hub を介してデータ連結を進めるなど、日本と類似の状況である可能性はあるが、活用状況が不明瞭のため引き続き調査を継続する。

F)マスタ情報の提供に関しては、今回のマスタ提供を機に、二次利用ポータル上での研究者等のマスタの共有が進むことを期待する。それとともに、研究班としても、また情報の提供を継続して行っていきたい。

本報告書で触れることなかった、⑥HIC を利用して解析可能なリサーチクエスチョンやその手法の例示、⑦研究者支援体制への協力についても進めており、次年度の報告書にて記載する。

## E．結論

本研究により、NDB/HIC の利活用拡大にむけてデータセット拡充の必要性を示すとともに、操作性改善案や標準化された SQL 処理、低コストなクラウド分析基盤等を提示した。今後各種調査・分析を進めるとともに、引き続き必要なデータ提供・支援を継続していく予定である。

## F．健康危険情報　なし

## G．研究発表

１.論文発表

1)　松居宏樹. 匿名医療保険等関連情報データベースの利用経験および今後の期待と課題. 統計. 2025；76（8）：14-22.

２．学会発表

1)　牧戸 香詠子、明神大也、森由希子　村松圭司　我が国およびフィンランドの医療データベースの最近の動向　第 84 回日本公衆衛生学会総会　2025 年 10 月 30 日

## 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

参考

Fukuyama K, Mori Y, Ueshima H, et al. Medical resource usage for COVID-19 evaluated using the National Database of Health Insurance Claims and Specific Health Checkups of Japan. PLoS One. 2024 May 13;19(5):e0303493.