別添3

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に関する研究
統括研究報告書

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	今村　知明	奈良県立医科大学公衆衛生学講座　教授
研究分担者	野田　龍也	奈良県立医科大学公衆衛生学講座　准教授
研究分担者	康永　秀生	東京大学大学院医学研究科　教授
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
研究分担者	中山　健夫	京都大学大学院医学研究科　教授

背景：近年、社会情勢や人口構造の変化に伴い。健康・医療・介護分野のビッグデータ解析の必要性が高まっている。匿名医療保険等関連情報データベース（NDB）及び介護保険総合データベース（介護DB）のデータについても利活用が期待されているが、データの容量が膨大であること、データの構造が複雑であること等から、データの構造や操作を知悉した研究者でなければ解析を行うのは容易ではない。NDBや介護DB等の大規模データを簡便に分析できるようになれば、健康・医療・介護ビッグデータの利活用推進に貢献することが期待される。

目的：「健康・医療・介護分野の大規模データの利活用を推進する」ことを目的に、NDB、および介護DB、さらに令和2年から提供が開始されている医療・介護の連結情報から、より利用者にとって使いやすいデータセットを開発する。また、医療・介護データ等の解析基盤（HIC）の試行利用を行い、安全性、操作性の検証を行う。

方法：本研究では、利用者によるデータセット設計を補助するために「①既存の大規模データベース（NDB・介護DB）のデータ構造の評価」を実施する。また、利用者にとって使いやすいデータセットの設計のために「②長期追跡性に優れた個人名寄せIDの検討」および「③臨床系研究者でも利用可能なデータセットの開発」について検討を行う。②および③については、研究の一部をHIC上で実施し、④HICの安全性と操作性について検討を行う。2023年度においては

① については、介護DBのデータ項目、データの格納状況について集計を実施し、結果を公表した。

② については、NDBに含まれる個人単位被保険者番号に基づく個人名寄せID（id5）が付与されたデータの長期追跡性について検討及び評価を行った。さらに、id5が付与される前のデータについても、既存のidとのより精緻なデータ連結の可能性について検討を行った。

③ については、今年度はNDBと介護DBの連結データの分析練習に使用可能なサンプルデータセットの仕様検討を行った。

④ については、実際のHICの利用を実施し、オンサイトリサーチセンターでのNDBデータ利用との比較検討を行った。

結果：

① 介護DBのデータについて網羅的な調査を行い、調査結果として作成したコードブックを学会およびホームページにおいて公表した。

② id5の設定状況を把握したとともに、id0およびid5の長期追跡の妥当性評価を行い、id0とid5を組み合わせた新たなid0nの開発に踏み込んだ。

③ NDBと介護DBの連結データの分析練習に使用可能なサンプルデータセットの開発を行った。

④ HIC利用における課題について、オンサイトリサーチセンター利用との比較およびHICにおいて比較的大規模な特別抽出データを取り扱う際の課題についての検討を行った。

結論： NDBにおける長期追跡可能なidに関する評価を行った。またNDB・介護DB連結データの練習用データセットを開発した。さらにHICにおけるデータ分析に関する課題抽出および解決法の検討を行った。これらの成果は今後利用者支援の一環となることが期待される。

Ａ．研究目的

本研究は「健康・医療・介護分野の大規模データの利活用を推進する」ことを目的に、匿名医療保険等関連情報データベース（NDB）データ及び介護保険総合データベース（介護DB）データ、さらに令和2年から提供が開始されている医療・介護の連結情報から、より利用者にとって使いやすいデータセットの開発を目指すものである。

昨今のコロナ禍に伴う社会状況や、近年の急速な少子高齢化を背景とする社会構造の変化に伴い、より適正な医療を提供するための基礎資料として、健康・医療・介護分野のビッグデータ解析の必要性が高まっている。実際NDB及び介護DBの情報はすでに医療・介護それぞれの分野での政策利用や研究利用が開始されており、提供件数は徐々に増加している。一方データの容量が膨大であること、データの構造が複雑であること等から、これらデータの構造や操作を知悉した研究者でなければ解析を行うのは容易ではなく、医療者等いわゆる臨床系の研究者だけでは取り扱いが困難なデータとなっている。一方で臨床系の研究者は、日々医療介護の現場で医療を担っており、適正な医療の提供・実施にあたっての課題にも直面していることから、こうした研究者が自らNDBや介護DB等の大規模データを簡便に分析できるようになれば、それら課題の解決に貢献することが期待される。

加えて、NDBは令和2年10月から介護DB及び令和4年4月からDPCデータベースとの連結が可能となり、さらに「医療・介護データ等の解析基盤（HIC）」が稼働されていることから、将来的には現在よりも利用の利便性が向上する見込みである。NDBデータを広く研究者が利用できるようにするためには、より簡便にデータ分析できる環境が整備される必要がある。

以上のような背景から、本研究班では、

①既存の大規模データベースの（NDB・介護DB）のデータ構造の評価
②長期追跡性に優れた個人名寄せIDの検討
③臨床系研究者でも利用可能なデータセットの開発

の3つの課題について検討を行うことを目的とした。また、②および③の一部をHIC上で実施（試行利用）することにより、④**HICの安全性及び操作性の検証**を行った。

Ｂ．研究方法（詳細については各分担研究報告書に記載）

本研究は、NDB と介護 DBさらに医療・介護の連結情報の利活用の推進を目的に、より多くの研究者が利用可能なデータセットの開発を行うものである。このため前述の3つの課題について令和5年度は以下のような方法で検討を行った。

①既存の大規模データベースの（NDB・介護DB）のデータ構造の評価

介護 DBデータを用いて、データ構造、テーブル構造、データ形式等の評価を行い、集計した情報について介護DBコードブックを作成し、公表した。

②長期追跡性に優れた個人名寄せIDの検討

NDBに含まれる個人単位被保険者番号に基づく個人名寄せID（id5）が付与されたデータの長期追跡性について検討及び評価を行った。

③臨床系研究者でも利用可能なデータセットの開発

NDBと介護DBはそれぞれ特有のデータ項目や構造があるため、連結データの分析にはそれらを理解した上での操作が必要となる。連結データ分析の練習を想定したサンプルデータセットの仕様検討を行い、作成した

⑤　HICの安全性及び操作性の検証

オンサイトリサーチセンターとHICにおけるNDBデータ利用について比較検討を行った。さらに特別抽出データを分析する上での課題について検討を行った。

Ｃ．研究結果（詳細については各分担研究報告書に記載）

①介護DBデータのコードブックの作成を行った。NDBとの連結データの利活用に資することを念頭に、NDBデータのコードブックの仕様をひな型とし、各項目の要約統計量（最大値、最小値、平均値、標準偏差など）による粗集計及び項目値のサンプル（最大値から上位50位までの項目値の実例）を算出する方針とし、集計可能なデータ項目、データ内容の精査を行い、実際の集計を行った。集計結果は第82回公衆衛生学会で報告し、京都大学医学部附属病院医療情報企画部ホームページ（https://medinfo.kuhp.kyoto-u.ac.jp/document/kaigo-db-codebook/）に公開した。

②id5の設定状況を調査した。全レセプトに含まれる有効なid5の割合はDPCレセプトでは95%程度、医科入院レセプトでは92%程度、医科入院外レセプトでは97%程度であった。またid0とid5から新たにid0nを作成した。2022年4月に存在したid0,id0n,id5に対し、2023年3月まで存在するidを追跡した結果、id0,id0n,id5いずれも、9ヶ月後で約90%の捕捉率を示し、id0nが最も捕捉率が高かった。

③。NDBと介護DBの連結データの分析練習に使用可能なサンプルデータセットの開発を行った。また、ユースケースの検討を行った。

④2023年2月からHICの試行利用を開始し、データの解析環境への取り込み、データベース構築を開始した。HICの構造およびデータアクセスとセキュリティ、解析環境の構築と大規模データの取扱いについて検討を行った。また、オンサイトリサーチセンターでの利用との比較についても検討を行った。

Ｄ．考察

介護DBに格納されているデータに関する基礎データを集計し、コードブックを作成した。NDBデータと同様に介護DBデータについても、コードブックを作成、公開することにより、今後介護DBデータに関する利用者の理解が深まり、データ利活用の一助となることが期待される。

長期追跡可能なIDの検討においては、NDBデータにおいてはすでにid0を用いた1患者1データ化の手法が開発されており、NDBデータ内における追跡調査・コホート調査が可能となっている（2021年度報告書にて報告）。今年度はid5の設定状況を把握したとともに、id0およびid5の長期追跡の妥当性評価を行い、id0とid5を組み合わせた新たなid0nの開発に踏み込んだ。この取り組みにより、すでに蓄積されている膨大な過去データとの連結解析が可能になり、様々な課題に対するデータの有効利用が期待される。

データセットの検討においては、研究者のニーズ

に対応できるデータセット作成に必要な条件を同定するために、2021年度、2022年度において実際にいくつかのリサーチクエスチョンについてNDBデータおよび自治体医療介護データを用いた分析を実施した（報告済み）。その結果、学術研究に資するようなデータセットの作成には、研究内容に即したデータセットの設計が必要であり、レディメイドのデータセットでは研究内容に制限がある可能性が示唆された。一方で、データセットの基本構造（エンティティ定義）等については共有できる可能性が改めて示唆された。今年度はいままでの検討結果をふまえて、NDB・介護DB連結データの分析練習ができるデータセットの開発を行った。簡便に使用可能なサンプルデータセットにより、連結データ利用に対するハードルが下がり、利用推進が期待される。

HICの試行利用においては、実際のデータを用いて、HIC上に分析用データベース構築を行い、HIC利用における課題検討を行った。また、HICとオンサイトリサーチセンターとの比較検討を行い、その特性の違いを説明した。

Ｅ．結論

①介護DBデータのコードブックは介護DBデータ及びNDB-介護DB連結データ利活用推進の一助となることが期待される。
②新規idであるid5のNDBにおける設定状況や長期追跡性を把握するとともにすでに開発されているid0とid5を組み合わせた新たなid0nの開発に踏み込んだ。
③「汎用性のあるデータセット」の一つとして、NNDB・介護DB連結データの練習用データセットを作成した。
④HICにおけるデータ利用の課題と対応策について検討を行った。

Ｆ．健康危険情報

なし

Ｇ．研究発表（別添4および各分担研究報告書に記載）

1. 論文発表　なし
2. 学会発表　あり

Ｈ．知的財産権の出願・登録状況

なし

1. 特許取得

なし

2. 実用新案登録

なし

3.その他

なし