

## NDB および自治体医療介護連結データを用いた臨床疫学研究

研究分担者 康永秀生 東京大学大学院医学系研究科臨床疫学・経済学 教授  
研究協力者 松居宏樹 東京大学大学院医学系研究科臨床疫学・経済学 助教

### 研究要旨

【目的】令和4年度は以下の臨床疫学研究等を実施した:(i) NDB データを用いた誘因イベント後の交感性眼炎の発生率の推計、(ii) 栃木県医療介護連結データを用いた要介護高齢者における心肺蘇生後の1年後の機能的転帰の推計。厚生労働省は2022年度、NDBデータのDaaS(Data as a Service)環境下での利用を開始した。本報告書ではその効率的な利用方法を検討した。

【方法】(i) 2012年から2019年の間に誘因イベント(inciting event)を経験した患者を対象とし、Cox 比例ハザード回帰分析により、交感性眼炎発症のリスク因子を検討した。(ii) 2014年6月から2018年2月までに心肺蘇生を受けた65歳以上患者を同定した。アウトカムは1年の死亡率および要介護度とした。また、NDBを利用した経験に基づき、DaaS環境下での利用方法を検討し、ORC経由でのDaaS利用(NDB ORC)と、HIC経由でのDaaS利用(NDB HIC)に焦点を当て、DaaSの構造、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した。

【結果】(i) 合計888,041件の誘因イベント(704,717人)が同定された。交感性眼炎の総症例数は263例であり、60ヶ月間の交感性眼炎の累積発生率は0.044%であった。40歳から49歳のグループは、最も高い0.104%の発症率を示した。

(ii) 65歳以上登録者594,092人の対象者のうち、5,389人(0.9%)が心肺蘇生術を受けた。要介護度なし、要支援1~2、要介護度1、要介護度2~3、要介護度4~5の患者における心肺蘇生後の1年死亡率はそれぞれ89.8%、89.2%、89.8%、89.9%であった。

NDBのDaaS環境下での利用方法に関して、現状を記述した。NDB ORCとNDB HICはアクセス経路と利用可能な環境が異なっている。前者においては、直接Redshiftを利用してNDBデータをハンドリングできる。しかし数千万症例規模のデータの解析が難しい場合があり、データ圧縮が必要となる。後者は、大規模データ解析は想定されておらず、工夫しないと解析が難しい。拡張機能をインストールし利用することで大規模データ解析は可能であるが、ハードルは高い。

【結論】(i) 今回の知見は、外傷後の交感性眼炎のリスクについて、また眼内手術を行う前に患者にカウンセリングを行う上で貴重なものとなる。(ii) 今回の知見は、患者、その家族、医療従事者が心肺蘇生に関する意思決定を共有することを支持

するエビデンスとなる。

ORC を用いた NDB の解析はかなり道筋がついたものの、HIC を用いた解析にはまだ課題が多く、改善策を検討すべきである。

## A. 研究目的

### 1. NDB 等を用いた臨床疫学研究

本研究班は「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、レセプト情報等データベース (NDB)、要介護認定情報・介護レセプト情報 (介護 DB)、および両者の連結情報から、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。

これらのデータは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。

研究全体の課題の一つとして、臨床系研究者でも利用可能なデータセットの開発が挙げられる。データセットの開発に先立ち、実際のデータ分析の実施を通じたデータ分析上の課題の整理が必要である。

本分担研究班 (東京大学) では、NDB を用いた臨床疫学研究、ならびに、研究分担者 (康永秀生) が併任する自治医科大学データサイエンスセンターの協力を得て利用している栃木県医療介護連結データを用いた臨床研究を進めている。2022 年度における成果を以下に 2 点報告する。

#### (i) 誘因イベント後の交感性眼炎の発生率

交感性眼炎は、眼の穿通性外傷または眼内手術後に対側眼に発生する肉芽腫性ぶどう膜炎である。フォークト-小柳-原田病と同じく、ぶどう膜のメラニン含有細胞に対する自己免疫反応が発生機序と考えられている。診断は臨床的に行われる。文献的に

は、交感性眼炎の約 80% の症例において、外傷または手術後 2~12 週以内にぶどう膜炎が生じた。

しかし本疾患は極めてまれであり、誘因となるイベント (眼外傷や眼内手術) 後の交感性眼炎の発生率は明らかとなっていない。本研究は NDB データを用いて、希少疾患である交感性眼炎の発生率を推計することを目的とした。

#### (ii) 要介護高齢者における心肺蘇生後の 1 年後の機能的転帰の推計

本研究は、栃木県医療介護レセプトデータベースを用いて、65 歳以上の要介護高齢者において、心肺蘇生後の 1 年間の機能的アウトカムを調査することを目的とした。

### 2. DaaS 環境下での NDB データの疫学研究への利用

厚生労働省は、NDB の解析環境を研究者向けに提供してきた。これにより、研究者は厚生労働省が保有する大量のデータを利用して、様々な研究や分析を行うことができるようになった。

2022 年度以前は、オンサイトリサーチセンター (ORC) において、解析端末を用意し、専用線を用いてオンプレミスのデータベースサーバーに接続し、そのうえで解析を行う解析システムを提供してきた。研究者は現地に足を運んで、安全にデータへアクセスができるようになり、解析環境を自前で用意する必要がなくなった。

2022 年度以降は NDB データの DaaS (Data as a Service) 環境下での利用が開始された。DaaS 環境では、データがクラウド上に保管され、リモート環境でのデータ利用が容易になる。

DaaS へのアクセス方法として、ORC 経由での利用と、セキュリティを確保した VPN 接続を介した利用 (HIC) が想定されている[図 1]。

ORC 経由での利用と、HIC 経由での利用は、いずれも AWS (Amazon Web Service) を利用している。

しかしながら、ORC 経由と HIC 経由では、利用可能な AWS サービスやインスタールされているソフトウェアが異なるため、適切な利用モデルの構築が必要である。研究者は、自分の研究目的や要件に応じて、最適な利用方法を選択する必要がある。

そこで、本報告書では過去の利用経験をもとに、NDB の DaaS 環境での利用方法を検討した。

## B. 研究方法

### 1. NDB 等を用いた臨床疫学研究

#### (i) 誘因イベント後の交感性眼炎の発生率

NDB を用いて後ろ向きコホート研究を実施した。対象者は 2012 年から 2019 年の間に誘因イベント(inciting event)を経験した患者とした。主なアウトカム評価項目は 60 ヶ月間以内の交感性眼炎の累積発生率とした。

交感性眼炎の発症は、交感性眼炎診

断の最初の日と定義した。

性別、年齢階級 (10 歳区切り)、一次または反復の誘因イベントにより層別化した上で、過去 1 年間の誘因事象の履歴 (誘因イベントなし、外傷なしの誘因イベント、外傷ありの誘因イベント) 別に交感性眼炎の累積発症率を Kaplan-Meier 法により算出した。Cox 回帰により調整ハザード比を推定した。

また、観察期間中に 1 回だけ誘因イベントがあった患者に母集団を限定し、誘因イベントの純粹な効果を推計した。

有意性の閾値は  $p < 0.05$  とした。すべての解析は R 3.0.1 (R Foundation for Statistical Computing, Vienna, Austria) を用いた。

#### (ii) 要介護高齢者における心肺蘇生後の 1 年後の機能的転帰の推計

栃木県医療介護レセプトデータベースに 2014 年 6 月から 2018 年 2 月までに登録された 65 歳以上の要介護高齢者のうち、心肺蘇生を受けた患者を特定した。主要アウトカムは、心肺蘇生後 1 年の死亡率および要介護度とした。

主要アウトカムは、心肺蘇生後 1 年の死亡率および要介護度とした。アウトカムは、1 日の推定介護時間の合計に基づき、心肺蘇生前に存在した介護ニーズによって、要介護度介護ニーズなし、要支援 1-2 および要介護

度 1 (推定介護時間 25-49 分)、要介護度 2-3 (50-89 分)、要介護度 4-5 (≧90 分) に層別化した。

## 2. DaaS 環境下での NDB データの疫学研究への利用

以下のプロジェクトを通じて、NDB を利用した経験に基づき、DaaS 利用環境での利用方法を説明する。

- ・ 「Covid19 流行の影響を利用した、医療のアウトカム研究」
  - ・ 「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」
- そのうえで、ORC 経由での DaaS 利用 (NDB ORC) と、HIC 経由での DaaS 利用 (NDB HIC) に焦点を当て、以下のポイントについて検討した。

### (i) DaaS の構造

どちらのアクセス方法も、NDB のメインデータベースに対する周辺システムとして構築がなされている。NDB メインシステムに関して考察を加えながら、それぞれの仕様を説明する。

### (ii) データアクセスとセキュリティ

どちらの利用方法でも、データアクセスとセキュリティが重要となる。ORC 経由と HIC 経由での DaaS 利用についてアクセス経路とセキュリティ上の対策を総括する。

### (iii) 解析環境の構築と大規模データの取扱い

どちらの利用方法でも、解析環境を

構築する際には、利用可能な AWS サービスやプリインストールされているソフトウェアを確認して、必要に応じて追加の設定やインストールを行う必要がある。各環境における解析環境整備の違いやライセンス運用に関して考察する。また、どちらの利用環境においても、大規模な症例データを解析する必要がある。環境の違いによる制約やそれに対する対処について検討した。

## C. 研究結果

### 1. NDB 等を用いた臨床疫学研究

#### (i) 誘因イベント後の交感性眼炎の発生率

合計 888,041 件の誘因イベント (704,717 人) が対象となった。交感性眼炎の総症例数は 263 例であり、60 ヶ月間の交感性眼炎の累積発生率は 0.044%であった。女性は交感性眼炎の発症と関連しなかった (ハザード比, 1.01; 95%信頼区間, 0.79-1.29;  $P=0.95$ )。40 歳から 49 歳のグループは、最も高い 0.104%の発症率を示した (80 歳グループ [0.041%] を対照とした調整ハザード比, 2.44; 95%信頼区間, 1.56-3.80;  $P < 0.001$ )。外傷を伴う、または伴わない反復した誘因事象は、一次誘因イベント (0.036%) よりも高い交感性眼炎の発生率 (それぞれ 0.469%、0.072%) を示した (ハザード比 11.68 [7.74-17.64] および 2.21 [1.59-3.07])。硝子体手術後の交感性眼炎の発生率は外傷後よりもはるかに低く (0.016%対 0.073%)、強膜バッ

クリング後の発生率はさらに低かった。

### (ii) 要介護高齢者における心肺蘇生後の1年後の機能的転帰の推計

594,092 人の対象者のうち、5,389 人 (0.9%) が心肺蘇生術を受けた。要介護度なし、要支援 1~2、要介護度 1 または要介護度 2~3、要介護度 4~5 の患者における心肺蘇生後の 1 年死亡率はそれぞれ 89.8% (n=2,207/2,457)、89.2% (n=736/825)、89.8% (n=930/1,036)、89.9% (n=963/1,071) であった。生存者のうち、ほとんどの患者は心肺蘇生前と 1 年後の要介護度に変化がなかった。潜在的な交絡因子を調整した後、既存の機能障害および認知障害と 1 年後の転帰との間に有意な関連は見られなかった。

## 2. DaaS 環境下での NDB データの疫学研究への利用

### (i) DaaS の構造

#### NDB 本体のアーキテクチャ

NDB は AWS 上に構築されている。NDB に格納されるレセプト情報と特定健診情報は、もともとテキスト情報である。大量のテキスト情報を AWS 上で管理する場合、テキストファイルを AWS S3 に保存し

(<https://aws.amazon.com/jp/s3/>)、AWS Glue(<https://aws.amazon.com/jp/glue/>) を用いて ETL 処理したうえで、Redshift を用いて抽出・集計するアーキテクチャが一般的である。

(<https://aws.amazon.com/jp/redshift/>)

また、処理するデータの量に応じ、

#### Redshift Spectrum

([https://docs.aws.amazon.com/ja\\_jp/redshift/latest/dg/c-using-spectrum.html](https://docs.aws.amazon.com/ja_jp/redshift/latest/dg/c-using-spectrum.html)) を用いて Redshift クラスター上にデータを移動することなく Redshift から直接データにクエリをかけることができる。NDB も上記のアーキテクチャを踏襲している。

#### NDB ORC の設計

NDB ORC は上記の NDB 本体の周辺システムとして設計がなされている。まず、NDB ORC 利用者は、専用端末の設置された個室から VPN 回線を用いて、AWS 上の Amazon Workplace にアクセスする。

(<https://aws.amazon.com/jp/workspaces/>)

次に、Workplace から AWS EC2

(<https://aws.amazon.com/jp/ec2/>) に構築

された解析環境へとアクセスすることで、解析を開始する。厚生労働省は NDB が格納された Redshift において、ORC 向けのデータベースを作成し、ORC の解析環境からの接続を許可している。そのため、EC2 上に構築された解析環境から Redshift 上に格納された ORC 向けデータベースを直接ハンドリングして解析を実施する。

[図 2]

#### NDB HIC の設計

NDB HIC も構造としては、NDB ORC と接続元とアクセス可能なデータが異なる点を除き本質的には同じもの

である。NDB ORC 利用者は、利用者自身の端末を用いて、AWS 上の Amazon Workplace にアクセスする。次に、Workplace から AWS EC2 に構築された解析環境へとアクセスすることで、解析を開始する。厚生労働省は NDB が格納された Redshift から部分的データを抽出し、CSV ファイルとして S3 上に保存したうえで、HIC の解析環境からの接続を許可している。そのため、EC2 上に構築された解析環境から S3 に保存された CSV ファイルをもとに、データベースを構築し解析を実施する。

#### (ii) データアクセスとセキュリティ

NDB ORC も HIC もデータ本体は、AWS 上の隔離された VPC (<https://aws.amazon.com/jp/vpc/>) に構築されており、外部からアクセスすることはできない。つまり、ORC も HIC もセキュリティ面で考えた場合、そのアクセス経路のセキュリティをいかに担保するかが重要となる。NDB ORC は HIC に比べ、物理的に隔離されたセキュリティレベルの高い環境からのデータへのアクセスが可能である。これに対して、HIC は物理的に隔離された環境を整備するわけではないため、アクセス可能なデータが部分的なものとなっている。

#### (iii) 解析環境の構築と大規模データの取扱い

NDB ORC, NDB HIC ともに解析環境として、m5.4xlarge (vCPU:16,

memory:64GB, Strage: 3TB) か、m5.2xlarge (vCPU:8, memory:32GB, Strage: 1TB) が利用可能である。なお、「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」は試行利用の扱いであるため、特例として 16TB のストレージを利用している。NDB ORC, NDB HIC ともに OS は、Windows, Linux を、データベースソフトウェアとして PostgreSQL を、統計解析ソフトウェアとしては R を利用可能である。また、事前申請に基づき、統計解析ソフトウェアとして、SPSS, Stata を追加で利用可能である。それに加え、NDB ORC においては、コンテナソフトウェアとして Docker を利用可能である。

#### NDB ORC での解析環境の整備

NDB ORC においては、解析環境から Redshift を直接参照可能であるため、データ抽出・ハンドリング処理の大部分を Redshift に off load できる。そのため、Redshift 上で、必要となる対象症例の情報を取り出し、データを整形した上で、解析環境上の PostgreSQL へデータを取り込んだ上で、必要な統計解析を行っている。しかしながら、対象症例の規模と、抽出する情報の粒度によっては、PostgreSQL 上にデータを取り込んだ段階で、解析環境のストレージ上限である 3TB をうわまわる事例がある。こうした事象に対応するため、Docker を用いてオープンソースの PostgreSQL の拡張である Citus

database (<https://www.citusdata.com/>)を構築し、列指向圧縮をデータにかけられることを検討した。実験では、列指向圧縮テーブルの容量は無圧縮テーブルの1/3~1/6ほどとなり、アクセス速度も、列数が少ない場合などにおいて向上した。この工夫により、数千万人規模の症例情報を与えられた解析環境で解析することができるようになった。

### NDB HIC での解析環境の整備

NDB HIC においては、解析環境から Redshift を直接参照できず、大規模なデータのハンドリングを行うための設計がなされていない。そのため、対象症例の規模を絞ったデータを予め抽出し S3 に保存し、そのデータを用いて統計解析を行う形が、想定される利用シーンとなっている。なお、「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」においては、試行利用の扱いであるため 10TB を超える大規模なデータを抽出している。HIC を用いて大規模なデータを処理するためには、いくつか環境に拡張機能のインストールが必要となる。NDB HIC では、EC2 の管理者権限がユーザーに付与されているため、必要となるソフトウェアのインストールを実施できる。

以下は、いくつかの拡張機能を用いた大規模データ処理の実施方法である。

- ・ Windows 環境の場合

- ・ 例えば Microsoft SQLSERVER 等の有料データベースソフトウェアを利用すること推奨する。これは、Windows 環境にインストールされた postgresql サーバーでは、いくつかの拡張機能が動作せず、大規模データのハンドリングが困難であるためである。
- ・ Linux 環境が用意できる場合、以下に示す適切な拡張機能のインストールが望ましい。
  - S3 上に保存されたデータを直接 PostgreSQL へと読み込むためには、PostgreSQL 拡張である、s3csv\_fdw のインストールが望ましい。
  - PostgreSQL において、テーブルデータの圧縮を行うことが望ましい。NDB ORC で実績のある Citus database のインストールが望ましい。

Docker 環境を立ち上げたうえで解析環境を整備することが可能であるならば、そうすべきである。

## D. 考察

### 1. NDB 等を用いた臨床疫学研究

#### (i) 誘因イベント後の交感性眼炎の発生率

60 ヶ月間の交感性眼炎の累積発生率は、最小でも 0.044%と推定された。繰り返される誘因イベント、特に外傷を伴う誘因イベントは、交感性眼炎発症のリスクを増加させた。外傷は硝子体手術に比べて交感性眼炎を



誘発する可能性が4~5倍であった。

## (ii) 要介護高齢者における心肺蘇生後の1年後の機能的転帰の推計

これまで要介護高齢者の心肺停止に対する心肺蘇生術の実施割合や予後について調べた研究はほとんどない。医療及び介護のレセプトデータを連結することにより、このような研究が可能となった。

本研究結果から、要介護度に関わらず、65歳以上の高齢者における心肺蘇生後の1年死亡率は約90%であった。生存者のうち、ほとんどの患者は心肺蘇生前と1年後の要介護度に変化がなかった。

## 2. DaaS 環境下での NDB データの疫学研究への利用

本報告書では、NDB の DaaS 環境での利用方法に関して、現状を記述した。NDB の DaaS 環境には NDB ORC と NDB HIC が存在し、アクセス経路と利用可能な環境が異なっている。NDB ORC においては、大規模な NDB データを、直接 Redshift を利用してハンドリングし解析することが可能である。また、解析環境の制約から数千万症例規模のデータを解析する事例において、解析が難しい場合があったが、データ圧縮を行うことで解析が可能となった。これに対し、NDB HIC は、大規模なデータを解析することは想定されておらず、適切な工夫を行わないと解析を行うことが難しいことが明かとなった。

いくつかの拡張機能をインストールし利用することで、大規模なデータのハンドリングを実施可能であるが、ハードルは高い。

今後、NDB HIC を本格運用するうえでは、いくつかのポイントを検討すべきである。第一に、HIC から Redshift へのアクセスを検討すべきである。現状、大規模なデータを対象とした解析が困難である。これは、Redshift へのアクセスを可能とすることで対処することができる。

次に、拡張機能のインストールをより柔軟に行えるように工夫すべきである。上述のように NDB HIC において、拡張機能のインストールが必要である。HIC 環境において Docker をインストールすることで対処することができる。

## E. 結論

### 1. NDB 等を用いた臨床疫学研究

#### (i) 誘因イベント後の交感性眼炎の発生率

今回の知見は、外傷後の交感性眼炎のリスクについて、また眼内手術を行う前に患者にカウンセリングを行う上で貴重なものとなるであろう。

#### (ii) 要介護高齢者における心肺蘇生後の1年後の機能的転帰の推計

今回の知見は、患者、その家族、医療従事者が心肺蘇生に関する意思決定を共有することを支持するエビデンスとなる。

本研究結果から、NDB データは比較的まれな疾患に対しても、そのデータ規模の大きさから、一定以上の症例数を確保でき、疾病の発生率、リスクファクターの探索研究が可能であることが分かった。その一方、疾病の重症度に関するデータが含まれないことから、治療の効果比較研究は困難であることが想定された。

医療介護連結データの利用にあたっては、それぞれのデータの特徴を理解した上で、研究の目的に沿うデータ項目をデータベースから選択し、分析モデルに適切に変数投入する必要がある点が明らかになった。特に介護データにおけるキー変数である要介護度のデータは時間依存性に変化するため、その取扱いに注意を要する。

## 2. DaaS 環境下での NDB データの疫学研究への利用

オンサイトリサーチセンターを用いた NDB の解析にはかなりの道筋があったが、HIC を用いた解析にはまだ課題が多い。今後の改善策を検討すべきである。

## F. 健康危険情報

なし

## G. 研究発表

### 1. 論文発表

- i. Hashimoto Y, Matsui H, Michihata N, Ishimaru M, Yasunaga H, Aihara M, Kaburaki T. Incidence of sympathetic ophthalmia after inciting events: a national database study in Japan. *Ophthalmology*. 2022;129(3):344-352.
- ii. Ohbe H, Nakajima M, Miyamoto Y, Shibahashi K, Matsui H, Yasunaga H, Sasabuchi Y. One-year functional outcomes after cardiopulmonary resuscitation for older adults with pre-existing long-term care needs. *Age and Aging*. 2023 in press

### 2. 学会発表

松居宏樹. NDB オンサイトセンターと臨床疫学研究基盤の構築～NDB ダミーシステムの作成～. 第5回 NDB ユーザー会. 2022年9月. 東京.

## H. 知的財産権の出願・登録状況

### 1. 特許取得

なし

### 2. 実用新案登録

なし

### 3. その他

なし

図 1 : 新 NDB システムとオンサイトセンター

## 新NDBシステムとオンサイトセンター

- NDBシステムのクラウド以降に伴い、2022年2月末日でオンサイトセンターは一端機能停止
- 2022年4月から、NDBシステムはAWSクラウド上に移行され、オンサイトセンターもクラウド上のシステムへの接続へ変更

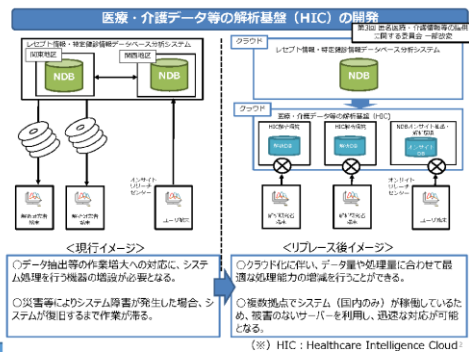


図 2 : オンサイトセンターでの NDB へのアクセス

## 新NDBシステムとオンサイトセンター

- 新システム下でのオンサイトセンターの利用

