令和7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））
「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」
分担研究報告書

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

研究分担者　松居宏樹 東京大学大学院医学系研究科准教授

研究要旨

本報告では、NDB、DPC、JMDC、DeSC などの大規模時系列医療データベースを疫学研究へ活用するため、標準化された SQL データハンドリングフレームワークを提示した。従来、SQL クエリ設計は研究者ごとの経験に依存し、再現性や保守性、レビュー容易性に課題があった。本研究では、「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。

A. 研究目的

大規模な時系列医療データベースの研究利用が急速に拡大している。わが国においては、匿名医療保険等関連情報データベース（NDB）、DPC 参加病院の退院患者情報を収録した DPC データベース、民間保険者の医療情報を収録した JMDC・DeSC データベース等、いずれも数テラバイト規模に達するデータが研究者に提供されるようになった（表 1）。これらのデータベースは、「誰が、いつ、どのような医療を受けたか」を記録した時系列情報の集積であり、疫学研究のきわめて有力なリソースである。

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。

本報告の目的は、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する。

B. 研究方法

B.1 大規模時系列医療データベース

わが国で研究利用される主要な大規模医療データベースを表 1 に示す。いずれのデータベースも、原則として診療の都度生成されるレセプト請求データを基盤としており、医薬品の処方・調剤、診療行為、傷病名、検査結果等の情報が時系列順に記録されている。

表 1. 主要な大規模医療データベース

データベース	収録対象	収録期間の目安	データ規模（概算）
DPC データベース	DPC 参加病院の入退院サマリ情報（全国）	2004 年～	年間約 6TB
JMDC データベース	一部保険者（社保）の医療レセプト・特定健診	約 17 年分	約 1.5TB
DeSC データベース	一部保険者（社保・国保・後期高齢）の医療レセプト・特定健診	約 8 年分	約 2.5TB
NDB データベース	全保険者の医療レセプト請求情報・特定健診	約 10 年分	約 10TB

B.2 NDB データベースの情報構造と時系列整理

NDB に収録される情報には、レセプト情報（医科・DPC・調剤・歯科）と特定健診情報が含まれる（表2）。近年では、死亡個票情報、匿名介護保険等関連情報データベースなどの公的データベース、次世代医療基盤法に基づく医療機関 EHR 情報等との連結も制度的に可能となり、活用可能な情報範囲は拡大している。

オンサイト・特別抽出データでは診療行為・医薬品等の実施年月日まで取得可能であり、日単位での詳細な時系列解析が可能である。一方、HIC 環境下の NDB-β では日付情報がマスクされ、月単位での解析に制限される。そのため、研究計画段階で利用環境を明確にし、利用可能な最小時系列粒度を前提にデータ整理を行う必要がある。

表 2. NDB データに含まれる主要情報区分と時系列解像度

医科レセプト

レコード種別	主な内容	時系列粒度
RE / HO / IR	レセプト共通ヘッダー（患者属性、診療年月、保険情報）	月単位
SY	傷病名、主傷病、修飾語、診断開始年月日	月単位 + 診断開始日
SI	診療行為、処置、検査、手術	日単位
IY	投薬情報（薬剤、用量、投与日数）	日単位
TO	手術・麻酔等詳細	日単位

DPC レセプト

レコード種別	主な内容	時系列粒度
RE	レセプト共通ヘッダー	月単位
SY / SB	傷病情報、診断開始年月日	月単位 + 診断開始日 / 入院単位
SI	出来高診療行為	日単位
IY	医薬品情報	日単位
TO	手術・処置	日単位
CD	DPC 包括評価関連コード	日単位

調剤レセプト

レコード種別	主な内容	時系列粒度
RE	レセプト共通ヘッダー	月単位
IY(SH)	調剤薬剤情報	日単位
CZ(SH)	調剤行為詳細	日単位
TO(SH)	特定加算・補足	日単位

歯科レセプト

レコード種別	主な内容	時系列粒度
RE	レセプト共通ヘッダー	月単位
SY / SB / SK	歯科傷病情報、開始日	月単位 + 診断開始日
SI	歯科診療行為	日単位
IY	歯科薬剤情報	日単位
TO	歯科処置・手術	日単位
SS	歯式情報	月単位または処置単位

B.2.1 NDB データの最小時系列粒度でのデータ整理

NDB データを最小時系列粒度で整理するためには、データ保存構造そのものを事前に設

計したうえで処理を進める必要がある。我々は「NDB Local PostgreSQL 最終出力テーブル仕様書」(添付1)に基づき、RE レコードを起点として患者 ID、レセプト ID、診療年月を整理する receheadertbl を作成している。

この設計により、「だれが」「いつ」「どうしたか」という疫学研究に必要な基本構造を保持しつつ、各診療行為情報を統合可能となる。さらに、SI、IY、TO レコードなど日単位情報を持つデータについては、実施年月日を付加することで、診療報酬請求コード単位で再整理を行う。

この一連の整理により、NDB データ全体を統一された最小時系列粒度で管理可能となる。HIC 環境では日付情報が制限されるため、月内代表日を設定して時系列表現を補完する。

B.3 大規模時系列医療データの特性

疫学研究において重要なのは、「だれが、いつ、どうしたか」を明確に記述することである。大規模医療データベースは本質的にその情報を蓄積しているが、研究利用には単純保存だけでは不十分であり、研究デザインに応じたデータハンドリングが必要となる。

したがって、大規模時系列医療データの活用には、時系列構造を維持しながら研究目的に即して再整理する体系的プロセスが求められる。

B.4 疫学的時系列構築の概念的フレームワーク

疫学研究デザインに対応した時系列構築には、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の 4 要素が必要となる。

解析単位とは、解析テーブル上の 1 行に対応する観察対象であり、個人単位、入院単位、手術単位、ICU 滞在単位など研究目的に応じて設定される。すべての解析単位は固有のタイムラインと Index date を持つ。

時間的アンカーとは、タイムライン上の基準日であり、Index date、イベント日、打ち切り日などが該当する。観察ウィンドウは、これらアンカーを起点として定義される時間区間であり、washout、baseline、exposure、follow-up など研究目的別に設定される。

また、診療コード・薬剤コード・病名コード等を研究カテゴリへ変換するため、マスタテーブルを整備する必要がある。これにより、コード管理の保守性と再利用性が確保される。

B.5 標準化された SQL ワークフロー

本研究では、概念的フレームワークを具体的な SQL 処理として実装するため、解析単位テーブル作成、可変アンカー付加、マスタ結合、ウィンドウベース集計、Long-to-Wide 変換という標準化ワークフローを採用した。

まず、Index date を基準とする解析単位テーブルを作成し、各観察単位に必要な可変アンカーを付加する。その後、マスタテーブルを用いて時系列データへカテゴリ情報を付与し、定義された観察ウィンドウごとに必要な情報を集計する。最終的に Long 形式の集計結果を Wide 形式へ変換し、解析可能な最終テーブルを構築する。

この構造化プロセスにより、再現性、可読性、保守性、計算効率を兼ね備えた標準的データハンドリングが可能となる。

C. 研究結果

C.1 標準化 SQL データハンドリングフレームワークの構築

本研究では、大規模時系列医療データを疫学研究へ効率的かつ再現性高く利用するため、

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の 4 要素を基盤とした標準化 SQL データハンドリングフレームワークを構築した。

このフレームワークにより、研究デザインの概念図（デザインダイアグラム）と SQL クエリ構造を一対一で対応させることが可能となった。すなわち、「だれが」「いつ」「どうしたか」という疫学研究の本質的構造をそのままデータベース設計とクエリへ反映できるようになった。

また、解析単位テーブルを起点として、個々の研究デザインに応じた可変アンカーを付加し、必要な患者・必要な観察期間のみを対象とした時系列データ抽出が可能となった。その結果、不必要なフルスキャンを回避しつつ、研究目的に即した効率的なデータ抽出が実現した。

C.2 SQL ワークフローの標準化

提案フレームワークに基づき、以下の一連の標準化ワークフローが整理された。

まず、Index date を持つ解析単位テーブルを作成し、研究ごとに異なる観察開始日・終了日・イベント日などの可変アンカーを付加した。次に、マスタテーブルを用いて時系列データへカテゴリ情報を付与し、観察ウィンドウごとの曝露・共変量・アウトカム情報を抽出した。さらに、Long 形式で集計した情報を Wide 形式へ変換することで、統計解析へ直接投入可能な最終解析テーブルを構築した。

この構造により、複雑な時系列データ処理が明確なモジュール単位に分割され、各工程の独立性と保守性が大きく向上した。

C.3 再現性およびレビュー容易性の向上

本フレームワーク導入により、SQL クエリの再現性は大幅に向上した。従来は担当者依存であったクエリ設計が、標準化された設計思想に基づいて構築されるため、研究者間での引継ぎや第三者レビューが容易となった。

特に、デザインダイアグラムとクエリ設計を対応させることで、コードレビュー時に研究デザインから逸脱した処理や不適切なウィンドウ設定を発見しやすくなった。これにより、ブラックボックス化した長大な SQL クエリの問題が大幅に軽減された。

C.4 計算効率および保守性の改善

解析単位テーブルを中心とした JOIN 構造により、必要最小限のデータ抽出が可能となり、大規模データベースに対する計算効率が改善した。さらに、コードのハードコーディングを排除し、マスタテーブル管理へ統一したことで、コード体系改定時や研究条件変更時の修正範囲が限定され、保守性が向上した。

加えて、中間テーブルの活用や CLUSTERED COLUMNSTORE INDEX の導入を前提とすることで、数 TB 規模のデータベースに対しても現実的な計算時間で運用可能な設計となった。

C.5 横展開可能性

本フレームワークは NDB データベースのみならず、DPC、JMDC、DeSC 等の他の大規模時系列医療データベースにも適用可能であることが確認された。各データベース固有のコード体系やテーブル構造の違いは存在するものの、解析単位、アンカー、観察ウィンドウ、マスタという基本構造は共通しており、マスタテーブルおよび一部テーブル定義を差し替えることで同一フレームワークを横展開できる。

この結果、本研究で提案する標準化 SQL データハンドリングフレームワークは、時系列医療データ研究全般に共通する基盤技術として利用可能であることが示された。

D. 考察

D.1 本フレームワークの意義

本報告で提示した標準化 SQL データハンドリングフレームワークは、従来属人的であった大規模医療データ解析を、疫学研究デザインに基づく体系的な設計へ転換する点に大きな意義がある。

最大の強みは、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である。従来の SQL 設計では、担当者個人の経験や記述習慣に依存しやすく、同一研究テーマであってもクエリ構造が大きく異なり、再現性やレビュー容易性に課題があった。本フレームワークでは、「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」という共通概念を用いることで、研究目的とクエリ構造の整合性が担保される。

さらに、解析単位テーブルを中心とするモジュール型構造は、可読性・保守性を大幅に向上させる。各工程が独立して管理されることで、研究条件変更時にも全体を書き換える必要がなく、限定的修正で対応可能となる。これは長期的な研究基盤整備において重要である。

D.2 再現性・教育的価値

標準化された構造は、単に個別研究の効率化にとどまらず、研究教育にも有用である。大規模医療データ解析は高度な SQL 技術を要求する一方、その教育体系は十分に整備されていないことが多い。

本フレームワークを用いることで、初学者であっても「どの順序で」「どの概念に基づき」データハンドリングを行うべきかが明確になる。これは教育コストを低減し、組織全体での解析品質向上に寄与する。

また、第三者レビューにおいても、研究デザインから SQL クエリへの変換過程が透明化されるため、疫学的方法論担当者、統計担当者、データエンジニア間の協働が容易になる。

D.3 パフォーマンスと実装面での重要性

数 TB 規模のデータベースを扱う現代の医療データ研究において、計算効率は研究実現可能性そのものに直結する。本フレームワークでは、解析単位テーブルを起点とする対象限定型 JOIN、マスタテーブル活用、中間テーブル管理、列指向インデックス活用などにより、パフォーマンス面でも合理的設計となっている。

これにより、不必要な全件スキャンや冗長なコピーを抑制し、計算資源消費を大幅に削減できる。したがって、本手法は単なる方法論的整備だけでなく、実務的運用基盤としても有用である。

D.4 限界

一方で、本フレームワークにはいくつかの限界も存在する。

第一に、データベースごとの固有仕様への対応は依然として必要である。NDB、DPC、JMDC、DeSC ではテーブル構造、コード体系、時間粒度が異なるため、個別マスタ整備やデータ理解は不可欠である。

第二に、コード体系の改定や制度変更に応じた継続的マスタ更新が必要となる。マスタテーブルを導入しても、その維持管理コストは残る。

第三に、研究デザインそのものが不適切であれば、SQL 標準化のみでは研究妥当性は保証されない。したがって、本フレームワークはあくまで「適切な研究デザインを高品質に実装する基盤」として位置づけられる。

D.5 将来的展望

本フレームワークは、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する。

研究デザインを構造化データとして保持できれば、将来的にはプロトコルから半自動的にSQL 生成を行う研究基盤も構築可能となる。その意味で、本研究は単なるクエリ標準化にとどまらず、大規模疫学研究インフラ整備の第一歩と位置づけられる。

E. 結論

本報告では、大規模時系列医療データを疫学研究に利用するための標準化 SQL データハンドリングフレームワークを提示した。

本フレームワークは、「だれが（解析単位）」「いつ（タイムライン）」「どうしたか（マスタ整備）」という疫学研究の本質的三要素を体系化し、解析単位テーブル作成、可変アンカー付加、時系列データ結合、観察ウィンドウ集計、Long-to-Wide 変換という一貫したワークフローとして整理した。

これにより、従来の属人的かつ非標準的であった大規模医療データ解析に対し、再現性、可読性、保守性、計算効率を兼ね備えた汎用的研究基盤を提供できる。

本手法は、NDB、DPC、JMDC、DeSC など多様な大規模時系列医療データベースへ適用可能であり、個別研究の品質向上のみならず、将来的な研究基盤標準化、教育体系化、半自動化研究環境整備にも寄与する。

大規模医療データ研究の発展に伴い、SQL データハンドリングの標準化は、今後ますます重要な基盤技術となると考えられる。

F. 健康危険情報

なし

G. 研究発表

1. 論文発表（著作）

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.

H. 知的財産権の出願・登録状況（予定を含む。）

なし