2018年度厚生労働科学研究費（統計情報総合研究事業）
患者調査等、各種機関統計調査におけるNDBデータの利用可能性に関する評価

各種基幹データ構成要素のNDB データとの代替可能性の検証：NDB分析用データセットの抽出・加工に関するバリデーション方法の開発

研究分担者　氏名　佐藤　大介
所属　国立保健医療科学院　保健医療経済評価研究センター　主任研究官

研究要旨

本研究では、厚生労働大臣から第三者提供を受けたレセプト情報等データベース（NDB）を加工して基幹データ構成要素の NDB データとの代替可能性の検証する際に生じる、中間生成物となる「分析用データセット」を生成するプロセスについて検証するためにダブルコーディングによるバリデーション方法について検討を行い、「分析用データセット」から「集計結果表」の生成で齟齬になりうる留意事項や処理プロセスで統一すべき事項について整理する。

本研究で用いたバリデーション手法によってバリデーションの結果が完全一致していることを確認するとともに、本研究を通じて処理プロセス別に起こりえる留意事項や処理プロセスで統一すべき事項について考察を行った。

同月での転院による複数回入院が含まれる場合の処理方法、２つ以上のレコードを結合する処理プロセスおよび結合方式の記載方法、医科レセプトと DPC レセプトによって抽出条件が異なる場合の取り扱い、欠損値の取り扱い等、解析計画書の正確な記述で誤った結果出力を防ぐことが可能となることが示唆された。

また、NDB はデータサイズが大きいため、まずは１万人サンプリング等の小さいデータベースを用いて解析を実施し、バリデーションが完全一致していることを確認したのち、最終解析を実行する方法が望ましい。

今後、各種基幹データ構成要素の NDB データとの代替可能性の検証の解析手順やバリデーション方法を明確にすることで、さらなる解析精度の向上が期待される。

Ａ．研究目的

本研究では、厚生労働大臣が保有し、第三者提供によって厚生労働省保険局から提供を受けたレセプト情報等データベース（NDB）を加工して中間生成物である「分析用データセット」を生成するプロセスについて、加工処理プロセスを検証するためにダブルコーディングによるバリデーション方法について検討する。加えて「分析用データセット」から「集計結果表」の生成で齟齬になりうる留意事項や処理プロセスで統一すべき事項について整理する。

Ｂ．研究方法

1.データソース

第三者提供によって厚生労働省保険局から提供を受けたレセプト情報等データベースの DPC レセプト、医科レセプト、調剤レセプトを用いた。

2.解析対象

（１）対象疾患

非弁膜症性心房細動：（I48.x）および心不全（I11.0、I50.0、I50.1、I50.9）を有する患者における医療費の集計（疑い病名を除く）

（２）選択基準

主傷病･副傷病に関わらず、非弁膜症性心房細動：（I48.x）、心不全（I11.0、I50.0、I50.1、I50.9）の傷病名（疑い病名は除く。）のいずれかが出現するレセプトを有する患者の全てのDPCレセプト、医科レセプト（入院・入院外）および当該医療機関と同一の医療機関コードからふり出された処方せんを調剤した薬局の調剤レセプト。

（３）除外基準

選択基準で指定した傷病名を有しないDPCレセプト、医科レセプト（入院・入院外）レセプトおよび当該医療機関と同一の医療機関コードからふり出された処方せんを調剤した薬局の調剤レセプトは非関連医療費と考えて除外する。

3.「分析用データセット」生成の処理プロセス

厚生労働省保険局から提供を受けたレセプト情報等データベース（NDB）から「解析用データセットテーブル」を生成する方法は、福田治久[1,2]の先行研究成果を用いることとし、「分析用データセット」を生成する処理プロセスは2.解析対象に基づき以下のステップで行う。

(0) 共通処理）

0.1　データベース管理システムにスキーマを作成する

0.2　レセプトに格納されているレコード情報別にテーブルを定義する

0.3　対象傷病名マスターを読み込む

0.4　フラグ傷病名マスターを読み込む

0.5　フラグ診療行為マスターを読み込む

0.6　フラグ医薬品マスターを読み込む

(1)「傷病名レコード情報」を「対象傷病名マスター」で絞り込んで「最古の入院IDのマスター」（患者ID、入院ID）を作成する。入院IDがNULL、疑いフラグ=1は除外する。

(2)「入院情報レコード」を「最古の入院IDのマスター」で絞り込んで「解析対象入院IDマスター」（入院ID、患者ID、最終診療日）を作成する。最終診療日がNULL（退院日が不明）は除外する。

(3)「傷病名レコード」を「解析対象入院IDマスター」「フラグ傷病名マスター」で絞り込んで「傷病名フラグマスター」を作成する。疑いフラグ=1は除外する。

(4)「傷病名レコード」を「解析対象入院IDマスター」「フラグ傷病名マスター」で絞り込んで「傷病名フラグ（主傷病に限定）マスター」を作成する。疑いフラグ=1は除外する。

(5)「診療行為レコード」を「解析対象入院IDマスター」「フラグ診療行為マスター」で絞り込んで「診療行為フラグマスター」を作成する。

(6)「医薬品レコード」を「解析対象入院IDマスター」「フラグ医薬品マスター」で絞り込んで「医薬品フラグマスター」を作成する。

(7)「診断群分類情報レコード」を「解析対象入院IDマスター」で絞り込んで「今回入院年月日マスター」（レセプトID、今回入院年月日）を作成する

(8)「レセプト情報レコード」を「解析対象入院IDマスター」で絞り込んで、「今回入院年月日マスター」および各フラグマスターを連結して解析対象入院期間データを作成する（レセプトID、レセプト総括区分、患者SID1、患者SID2、患者ID、診療年月、診療時年齢区分、男女区分、入院年月日、今回入院年月日、合計点数、診療実日数、入院ID、レセプト区分、集計

区分、医薬品フラグ、診療行為フラグ、傷病名フラグ、主傷病名フラグ）。対象は医科レセプトと DPC レセプトとする。DPC レセプトのレセプト総括区分=1 は除外する。

(9)「解析対象入院 ID マスター」から「退院後診療年月マスター」（患者 ID、診療年月、集計区分(2/3/4)）を作成する。

(10)「傷病名レコード」を「退院後診療年月マスター」「フラグ傷病名マスター」で絞り込んで「退院後傷病名フラグマスター」を作成する。疑いフラグ=1 または「解析対象入院 ID マスター」に含まれる入院 ID を持つレコードを除外する。

(11)「傷病名レコード」を「退院後診療年月マスター」「フラグ傷病名マスター」で絞り込んで「退院後傷病名フラグ（主傷病に限定）マスター」を作成する。疑いフラグ=1 または「解析対象入院 ID マスター」に含まれる入院 ID を持つレコードは除外する。

(12)「診療行為レコード」を「退院後診療年月マスター」「フラグ診療行為マスター」で絞り込んで「退院後診療行為フラグマスター」を作成する。「解析対象入院 ID マスター」に含まれる入院 ID を持つレコードは除外する。

(13)「医薬品レコード」を「退院後診療年月マスター」「フラグ医薬品マスター」で絞り込んで「退院後医薬品フラグマスター」を作成する。「解析対象入院 ID マスター」に含まれる入院 ID を持つレコードは除外する。

(14)「診断群分類レコード」を「退院後診療年月マスター」で絞り込んで「退院後今回入院年月日マスター」（レセプト ID、今回入院年月日）を作成する

(15)「レセプト情報レコード」を「退院後診療年月マスター」で絞り込んで、「退院後今回入院年月日マスター」および各退院後フラグマスターを連結して解析対象退院後データを作成する（レセプト ID 、レセプト総括区分、患者 SID1、患者 SID2 、患者 ID 、診療年月、診療時年齢区分、男女区分、入院年月日、今回入院年月日、合計点数、診療実日数、入院 ID、レセプト区分、集計区分(2/3/4) 、医薬品フラグ、診療行為フラグ、傷病名フラグ、主傷病名フラグ）。「解析対象入院 ID マスター」に含まれる入院 ID を持つレコードまたは DPC レセプトのレセプト総括区分=1 は除外する。

(16)解析対象入院期間データ、解析対象退院後データを連結して疾病費用分析用データセットを作成する。（レセプト ID 、レセプト総括区分、患者 SID1、患者 SID2 、患者 ID 、診療年月、診療時年齢区分、男女区分、入院年月日、今回入院年月日、合計点数、診療実日数、入院 ID、レセプト区分、集計区分(2/3/4) 、医薬品フラグ、診療行為フラグ、傷病名フラグ、主傷病名フラグ）

４．バリデーション方法

（１）分析担当者が「解析用データベース」から「疾病費用分析用データセット」を生成する処理プロセスを処理プロセス計画書（３．処理プロセス計画書参照）に記載する。

（２）作成した処理プロセスのアルゴリズムを分析担当者と分析担当 SE が確認を行い、疑問点や留意事項について合意した内容を処理プロセス計画書に加筆修正し、処理プロセスを確定する。（次節　処理プロセス計画　参照）

（３）分析担当者と分析担当 SE がそれぞれ独立して処理プロセス計画書を実行するプログラムを作成する。プログラム言語は SQL 言語（PostgreSQL）に統一する。
（４）分析担当者と分析担当 SE がそれぞれ作成したプログラムを実行し、生成された「疾病費用分析用データセット」が完全一致するかどうかを比較する。比較方法は Linux コマンドの”diff”を使用する。
（５）完全一致しない場合は処理プロセス計画書のプロセス毎に生成する中間テーブルを比較し、分析担当者と分析担当 SE によるコードレビューを行い、処理プロセス計画書の記載内容やプログラムに不備等の修正を行う。修正後に再度生成した「疾病費用分析用データセット」が完全一致するかどうか比較する。（完全一致するまで繰り返す）
（６）分析担当者と分析担当 SE が生成した「疾病費用分析用データセット」が完全一致していることを確認した後、diff コマンドの出力結果画面のハードコピー（Print Screen）を保存する。

Ｃ．研究結果

「解析用データベース」から「疾病費用分析用データセット」を生成する処理プロセスについて、分析担当者と分析担当 SE が生成した「疾病費用分析用データセット」が完全一致していることを確認した。（別図：diff コマンドの出力結果画面のハードコピー（Print Screen）参照）

Ｄ．考察

医療費用を解析するにあたり、基礎データとなる「疾病費用分析用データセット」および「集計結果表」の生成について、加工処理プロセスを検証するためにダブルコーディングによるバリデーションを実施した結果、プログラムコードや処理プロセス手順についての検討・修正を経て、結果が完全一致していることを確認した（別図参照）。本研究で実施にあたっては、処理プロセスのステップ別に起こりえる留意事項や処理プロセスで統一すべき事項について考察する。

第一に、解析対象マスターとする「最古の入院 ID のマスター」を作成する際、「最古の入院」は最古の診療年月ではなく、入院 ID の最小値とすることを明記する。（理由：最古の診療年月で絞り込むと、同月での転院による複数回入院が含まれる。そのため当該マスターが一意の患者とならない。その結果、最終データセットを作成すると件数が倍増する。）「傷病名フラグマスター」「主傷病名フラグマスター」「診療行為フラグマスター」「医薬品フラグマスター」は３つのテーブルを結合して生成するが、段階的に処理する方法と１つに纏めて処理する方法がある。３つ以上のテーブルを結合する処理プロセスは可能な限り段階別に記載するのが望ましい。各処理プロセスで作成したマスターと「最古の入院 ID のマスター」と結合し一つのテーブルに集約する際、結合方式が INNER JOIN（絞り込み）か「最古の入院 ID のマスター」に揃える LEFT　JOIN（連結）かがわかるよう記載する必要がある。入院期間におけるレセプトを処理するプロセスでは、医科レセプトと DPC レセプトによって抽出条件が異なる場合は具体的に明記する。（例：DPC レセプトのレセプト総括区分の条件）また、退院後におけるレセプトを処理するプロセスでは、入院レセプトを除外することや、各処理プロセスで作成する中間テーブルの名称を明記する。テーブル名称は「」で括ることで処理プロセスの読みやすさを改善することができる。各処理プロセスで設定する条件について、NULL の取り扱いについて明記する。各マスターを作成するプロセスにおいて、レセプト ID 単位で一意となるテーブルを前提とする場合、データはレセプト ID が一意

（DISTINCT）であることを明記する。対象傷病名、対象医薬品、対象診療行為等の対象マスタを事前準備する際、フラグ項目に重複する傷病名コード、医薬品コード、診療行為コードが存在する場合、解析デザイン上必要か確認し、プログラムを作成する必要がある。

なおNDB解析用データベースから疾病費用分析用データセットを抽出・加工する処理時間は、各種レセプトやレコードのデータ量が大きいため、作業付加が大きい。そのためバリデーションでは、まずは1万人サンプリング等の小さいデータベースを用いて実施し、双方の処理結果が完全一致していることを確認したのち、10%または100%データを用いて最終確認する方法が望ましい。これはバリデーションだけでなく、公的分析プロセスにおけるNDB分析実施可能性の検討や分析枠組みが確定する前の解析において柔軟な対応を可能にすることが期待される方法である。

E．結論

本研究では、非弁膜症性心房細動：（I48.x）および心不全（I11.0、I50.0、I50.1、I50.9）を有する患者を対象に、国内におけるレセプトのデータベースの一例として厚生労働省から提供を受けたレセプト情報等データベース（NDB）を利用した疾病費用分析を実施するために「解析用データベース」から疾病費用分析の基礎となる「疾病費用分析用データセット」の生成プロセスおよび費用解析により算出した「集計結果表」について、加工処理プロセスを検証するためにダブルコーディングによるバリデーションを実施した。その結果、プログラムコードや処理プロセス手順についての検討・修正を経て、結果が完全一致していることを確認した。これらの結果に基づき、生成プロセスで齟齬になりうる留意事項や処理プロセスで統一すべき事項を整理した。

今後、各種基幹データ構成要素のNDBデータとの代替可能性の検証の解析手順やバリデーション方法を明確にすることで、さらなる解析精度の向上が期待される。

【参考文献】

・「保健医療科学」68巻2号 福田治久，佐藤大介，白岩健，福田敬 NDB解析用データセットテーブルの開発
・ 「保健医療科学」68巻2号 福田治久，佐藤大介，福田敬 レセプトデータを用いた医療費分析における診療報酬改定の補正方法

F．研究発表

1．論文発表

該当無し

2．学会発表

該当無し

G．知的財産権の出願・登録状況

（予定を含む）

1．特許取得

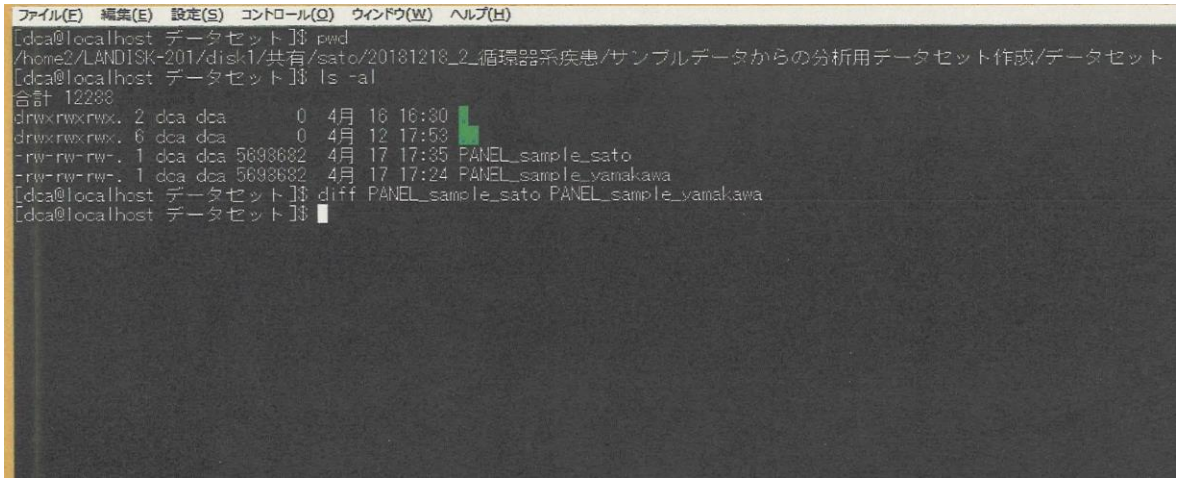
該当無し

2．実用新案登録

該当無し

3．その他

該当無し



（別図：diff コマンドの出力結果画面のハードコピー（Print Screen））