厚生労働科学研究費補助金（認知症政策研究事業）
分担研究報告書

KDB システム等を用いた自治体事業の質の評価に関する研究：「突合データ（CSV)」ファイルのデータベース開発

研究分担者　石崎達郎　東京都健康長寿医療センター研究所・研究部長
研究協力者　光武誠吾　東京都健康長寿医療センター研究所・研究員
研究協力者　深谷太郎　東京都健康長寿医療センター研究所・研究員
研究協力者　杉山美香　東京都健康長寿医療センター研究所・研究員

研究要旨

国保データベース（KDB）システムに格納されている要介護認定情報や介護サービスの利用状況、受療状況等に関する情報を活用することで、独居認知症高齢者等が地域での生活継続を可能とするためのサービス提供の検討や地域システムの質評価に有用な情報を創出することが可能となることから、令和４年度は特定自治体から提供を受けた KDB システム「突合データ（CSV)」を用いて研究所内にて独自にデータベースを開発し、KDB システムを用いた自治体事業の質の評価を可能とする基盤の構築を目的とした。KDB システム「突合データ（CSV)」に含まれる CSV 形式データの文字コード処理を施したのちに、マイクロソフト社 SQL Server を使って 12 種類のファイルごとにデータベースを構築した。データ提供元の自治体担当職員と「突合データ（CSV)」の提供方法について意見交換を行ったことに加え、CSV 形式データの文字コード処理方法について東京都国保連合会への照会を重ねながら、最終的にデータベースを構築することができた。今後は、独居認知症高齢者等が地域での生活継続を可能とするためのサービス提供や地域システムの質評価に有用な情報の抽出方法を検討する。

A. 研究目的

国保データベース（KDB）システムは、保険者によるデータヘルス計画の策定・実施等を支援するために、国民健康保険団体中央会（国保中央会）によって開発されたデータベースシステムである。このシステムには健康診査、医療・介護保険レセプトの各種データが突合・加工され、専用のシステムを介して、全国の市町村や保険者等に統計情報や個人の健康に関するデータが提供されている。

この KDB システムに格納されている要介護認定情報や介護サービスの利用状況、受療状況等に関する情報を活用することで、独居認知症高齢者等が地域での生活継続を可能とするためのサービス提供の検討や地域システムの質評価に有用な情報を創出することが可能となる。そこで令和４年度は、特定自治体から提供を受けた KDB システム「突合データ（CSV)」を用いて

研究所内で独自にデータベースを開発し、KDB システムを用いた自治体事業の質の評価を可能とする基盤構築を目的とする。

B. 研究方法

１．使用データ（ファイル）

東京都内の某自治体から提供を受けた KDB システムの「突合データ（CSV)」に含まれる以下のデータをデータベース開発に使用した。

１）KDB 被保険者台帳（国民健康保険、後期高齢者医療、介護保険）：KDB 個人番号、性別、氏名（削除)、住所（削除)、生年月日、郵便番号、通番、国保取得年月日、国保喪失年月日、後期取得年月日、後期喪失年月日、介護保険初回認定日、初回要介護度、直近の要介護度、国保喪失事由、後期喪失事由等

２）健診結果（国民健康保険・後期高齢者医療）：KDB 個人番号、保険者番号、測定値・検査値（身長、体重、BMI、血圧、血清脂質、肝機能、空腹時血糖、HbA1c、尿糖、尿たんぱく、貧血検査、尿酸、血清クレアチニン、eGFR 等)、服薬状況、既往歴、喫煙、質問票（標準的な質問票、後期高齢者の質問票）等

３）医療レセプト管理（国民健康保険・後期高齢者医療、医科・歯科）：KDB 個人番号、診療年月、受付番号、医療機関コード、レセプト種別、入外区分、診療実日数、入院年月日、決定点数、診断群分類番号、DPC 転帰区分コード、郡市区コード

４）医療傷病名（国民健康保険・後期高齢者医療、医科・歯科）：KDB 個人番号、審査年月、受付番号、医療機関コード、診療開始年月日、傷病名、傷病名コード、修飾語コード、疑い病名区分、ICD10 コード等

５）医療摘要（国民健康保険・後期高齢者医療、医科・歯科）：KDB 個人番号、診療年月、受付番号、適用コード（診療行為、薬剤等)、単位コード、数量、回数、点数、薬効分類等

６）医療最大医療資源傷病名コード別点数（国民健康保険・後期高齢者医療）：KDB 個人番号、審査年月、受付番号、医療機関コード、傷病名コード（ICD10)、診療開始年月日、合計点数等

７）介護給付基本実績（介護保険）：KDB 個人番号、審査年月、サービス提供年月、事業所番号、要介護状態区分コード、サービス種類コード、サービス点数、保険請求額等

８）介護給付実績明細（介護レコード種別ごとに 11 種類）：KDB 個人番号、審査年月、サービス提供年月、要介護状態区分コード、サービス種類コード、サービス項目コード、日数回数、サービス点数、適用情報、緊急時傷病名、往診日数、往診医療機関名、通院日数、通院医療機関名、緊急時治療管理点数、緊急時治療日数、サービス実日数、点数合計、保険請求額等

９）医療受診日等（国民健康保険・後期高齢者医療、医科・歯科）：KDB 個人番号、診療年月、受付番号、医療機関コード、レセプト種別区分、受診等区分（01 日～31 日）

１０）医療摘要回数（国民健康保険・後期高齢者医療、医科・歯科）：KDB 個人番号、診療年月、受付番号、医療機関コード、レセプト種別区分、診療識別、回数（01 日～31 日）等

１１）介護受給者認定情報（介護保険）

２．使用したパソコンソフト

テキストエディターは、2GB を超えるファイルにも対応可能なテキストエディターとして Emurasoft 社の EmEditor を使用した。リレーショナルデータベース（RDB）の構築にはマイクロソフト社 SQL Server を使用し、データベースは 2GB を超えるため有償版を購入した。

（倫理面への配慮）

東京都健康長寿医療センター研究倫理委員審査委員会の承認を得て実施した。

C. 研究結果

１．「突合データ（CSV）」の入手

「突合データ（CSV）」（KDB 被保険者台帳ファイルを除く）は自治体で使用されている KDB システム端末からダウンロードすることができないため、本研究でデータ提供を受けた東京都内某自治体が東京都国民健康保険団体連合会（東京都国保連）に「突合データ（CSV）」の全ファイルの提供を依頼することで、データを入手することができた。

「KDB 被保険者台帳」ファイルだけが氏名や住所等の個人を識別可能な情報が含まれ、それ以外の「突合データ（CSV）」のファイルには氏名や住所等の情報は含まれていない。その代わりに個人を識別する「KDB 個人番号」がすべてのファイルに含まれている。そこでこの「KDB 個人番号」を各ファイルの連結キー情報として使用することで、各ファイルに含まれる情報を統合した分析用ファイルを作成することが可能となる。

なお本研究でデータベースを構築する際、当該自治体の庁舎内で「KDB 被保険者台帳」ファイルに含まれる氏名、住所、電話番号等の削除処理を施した後に、この台帳データの提供を受けた。

「突合データ（CSV）」に含まれるデータは、2016 年 6 月から 2022 年 6 月までのデータで、被保険者台帳は 2022 年 12 月時点でのデータであった。

２．CSV ファイルの前処理

東京都国保連合会から提供された「突合データ（CSV）」に含まれるファイルは、その名の通りすべて CSV 形式で、文字コードは「UTF-16LE」が使用されていた。また、これらのファイルは、そのテキスト先頭にバイト順マーク（Byte Order Mark、BOM）がついていない「BOM なし」形式であったため、このままではテキストエディターやエクセル等でファイルを開いても、全角文字は文字化けしてしまい、データベース構築に移ることができなかった。文字コード処理方法についてデータ提供元の東京都国保連の KDB データ担当者に照会を重ねながら、最終的に「BOM なし」形式を「BOM あり」形式に変換する必要のあることが分かった。そこでデータベース構築前の処理として、すべてのファイルを「UTF-16LE（BOM なし）」から「UTF-16LE（BOM あり）」に変換した。

テキストエディターを使えば、手作業で一つ一つのファイルを「UTF-16LE（BOM あり）」へ変換して保存することは可能ではあるが、この作業の対象となるフ

ファイルは約 2000 種類あるため、一つ一つを手作業で処理することは現実的ではなかった。そこで、Windows のバッチファイルを独自に作成し、すべてのファイルを一括処理して「UTF-16LE（BOM なし）」から「UTF-16LE（BOM あり）」に変換した。具体的には、次の BOM をテキストファイルの先頭に付加するバッチファイルを作成した。

UTF-16LE の場合 0xFF 0xFE
（BOM の内容は、バイナリエディタを使用しないと目視できない）

CSV ファイルを「UTF-16LE（BOM あり）」に変換したことで、テキストエディターを使ってファイルの中身を目視で確認できるようになった。その過程で、医療費情報はデータ中に千単位でカンマが含まれていることがわかった。また、スペース等が含まれる列のデータは、データがダブルクォーテーション（"）で囲まれていることもわかった。そこでデータベースにファイルを取り込む前に、テキストエディターを使って、医療費のカンマとダブルクォーテーションを削除することで、データベースに取り込む CSV ファイルを完成させた。

３．ファイルレイアウトの変更

今回、当該自治体から提供を受けた突合データ CSV は、2016 年 6 月から 2022 年 6 月までの 73 か月間のデータであった。この間、ファイルレイアウトが変更されていたものがあったため、表 1 に示す通り、ファイルレイアウトの変更履歴を把握した後に、同じレイアウト毎にファイルを統合してから、データベースに組み入れた。

４．データベースの構築

マイクロソフト社 SQL Server によるデータベース構築は、次の二種類のクエリを使用した。まず、「CREATE TABLE」クエリを使用して各 CSV データを流し込むデータレイアウトを作成した。次に作成されたテーブルに CSV データを流し込むために、「BULK INSERT」クエリを使用した。

D. 考察

独居認知症高齢者等が地域での生活継続を可能とするためのサービス提供や地域システムの質評価に有用な情報を創出するために、本研究の対象自治体から提供を受けた KDB システムの「突合データ（CSV）」を用いて研究所内にて独自にデータベースを開発し、KDB システムを用いた自治体事業の質の評価を可能とする基盤を構築した。「突合データ（CSV）」（KDB 被保険者台帳ファイルを除く）は自治体で使用されている KDB システム端末からダウンロードすることができないため、本研究を実施した東京都内某自治体から東京都国保連に「突合データ（CSV）」全ファイルの提供を依頼し、東京都国保連から自治体に提供されたデータを我々が入手した。東京都国保連から「突合データ（CSV）」のデータレイアウト一覧も入手した。このレイアウト一覧には、突合データ CSV の各ファイルの情報の属性（9:数字、X:英数字、N:全角文字）と長さ（単位：バイト）が示されており、データベースに格納するテーブルを作成際は、このデ

ータレイアウト一覧が有用である。ただし実際のファイルの中には、データレイアウトに記されているバイト数よりも大きなバイト数を必要とするデータが存在していた。例えば、表２に示すように、健診結果ファイルには、「体重」や「検査値_血清クレアチニン」は、データレイアウトではデータ長は「4 バイト」（小数点を入れて 4 ケタ）と示されているが、体重 100 キロ超の者、透析患者で血清クレアチニンが 10 を超える者では「5 バイト」（小数点を入れて 5 ケタ）が必要となる（例：体重「100.3」、クレアチニン「10.24」）。このような外れ値を有するデータが含まれるファイル項目については、データレイアウトに記されているバイト数よりも大きな値を任意に再設定することで、「BULK INSERT」クエリをエラー無く完了させることができた。

E. 結論と今後の課題

東京都内の某自治体から提供を受けた KDB システム「突合データ（CSV)」を用いて、研究所内において独自に RDB を開発し、KDB システムデータを用いて、独居認知症高齢者等が地域での生活継続を可能とするための自治体事業の質の評価を可能とする基盤を構築した。「突合データ（CSV)」のデータレイアウトは全国共通ではあるが、CSV ファイルの文字コード（UTF-16LE）や BOM 付与対応等、国保連合会からのデータ提供時に確認・依頼すべきポイントが明らかとなった。来年度は開発されたデータベースを用いて、独居認知症高齢者等が地域での生活継続を可能とするためのサービス提供や地域システムの質評価に有用な情報の抽出方法を検討する。

F. 研究発表

1. 論文発表

該当なし

2. 学会発表

該当なし

G. 知的財産権の出願・登録状況

1. 特許取得

該当なし

2. 実用新案登録

該当なし

3.その他

該当なし

表１.KDB システム「突合データ（CSV)」に含まれるファイル種類とデータレイアウトの変更履歴

ファイル種類	保険種類、医科・歯科	入手データ期間（処理年月）とレイアウト変更の有無
KDB 被保険者台帳	国保・後期・介護保険	2022 年 6 月(全ての被保険者の履歴を含む)［ファイル数：3］
医療レセプト管理	国保・後期、医科・歯科	2015 年 6 月～2022 年 6 月（変更なし）［ファイル数：340］
医療傷病名	国保・後期、医科・歯科	2015 年 6 月～2022 年 6 月（変更なし）［ファイル数：340］
医療摘要	国保・後期、医科・歯科	2015 年 6 月～2022 年 6 月（変更なし）［ファイル数：340］
健診結果	国保・後期、医科・歯科	2015 年 6 月～2018 年 5 月[ファイル数:144]
	国保・後期、医科・歯科	2018 年 6 月～2020 年 5 月（「基本チェックリスト」追加）［ファイル数：96］
	国保・後期、医科・歯科	2020 年 6 月～2022 年 6 月（「後期高齢者の質問票」追加）［ファイル数：100］
医療受診日等	国保・後期、医科・歯科	2022 年 3 月～2022 年 6 月（変更なし）［ファイル数：16］
医療摘要回数	国保・後期、医科・歯科	2022 年 3 月～2022 年 6 月（変更なし）［ファイル数：16］
医療最大医療資源 ICD 別点数	国保・後期、医科のみ	2015 年 6 月～2018 年 5 月[ファイル数：72]
	国保・後期、医科のみ	2018 年 6 月～2022 年 6 月（「ICD-10 関連項目」追加）［ファイル数：98］
介護給付実績		2015 年 6 月～2022 年 6 月（変更なし）［ファイル数：85］
介護受給者認定情報		KDB 処理年月 2022 年 6 月[ファイル数：1]
介護給付実績明細	介護レコード種別「13 種類」でレイアウト異なる	2021 年 6 月～2022 年 6 月[ファイル数:143]
介護総合事業実績	介護レコード種別「3 種類」でレイアウト異なる	2021 年 6 月～2022 年 6 月[ファイル数：39]

表２.「突合データ（CSV)」のファイルレイアウト例：データ種類～健診結果（抜粋）

データ種類	②健診結果			
No.	項目名	属性	長さ	精度
1	KDB個人番号	X	11	0
2	保険者番号	X	8	0
3	年度	9	4	0
4	データ管理番号1	X	10	0
5	受診券整理番号	X	22	0
6	健診実施年月日	9	8	0
7	健診機関コード	X	10	0
8	検査値_身長	9	4	1
9	検査値_体重	9	4	1
10	検査値_BMI	9	4	1
11	検査値_内臓脂肪面積	9	4	1