健康医療データベースに於ける暗号化された識別子に基づく患者の追跡方法の検討と実験に基づく考察

佐藤 淳平† 山田 浩之† 合田 和生† 喜連川 優† 満武 巨裕‡

†東京大学 生産技術研究所 〒153-8505 東京都目黒区駒場 4-6-1
‡医療経済研究機構 〒105-0003 東京都港区西新橋 1-5-11 第 11 東洋海事ビル 2F
E-mail: †{jsato,hiroyuki,kgoda,kiture}@tkl.iis.u-tokyo.ac.jp, ‡mitsutake@ihep.jp

あらまし 政府や企業をはじめとして，大規模なデータが蓄積されるようになり久しい．データの所有者が十分な解析能力を有しているとは限らず，データの解析を第三者に依頼して実施することも多い．第三者へのデータ提供の際には法令や契約により，氏名や生年月日などの個人を特定可能な属性情報に対する匿名化が義務付けられている．データを解析する際にはこれらの匿名化された属性情報を，個人を識別するための識別子として使用することが多いものの，これらの暗号化された属性情報はライフイベントを契機として変化することがあり，即ち，識別子も変化することから，個人の追跡ができなくなる可能性がある．本論文では我が国最大規模の健康医療データベースであるレセプト情報・特定健診等情報（通称ナショナルデータベース）の 6 年分のデータセットを対象として，２種類の暗号化された個人識別子に基づく個人の追跡手法を提案し，その識別可能性の検証結果を示す．

キーワード 個人識別子, 暗号化

1. はじめに

近年，国家や企業に於いて，ログデータをはじめとした様々なデータを蓄積し，蓄積データの解析結果に基づく組織の運営方針の決定が行われ始めている．総務省のデータ流通量に関する調査によると，電子カルテデータや画像診断データなど機微性の高いデータのデータ量が増加していることが示されている[4]．しかしながら，データの膨大性のためデータの所有者自身が十分な解析を行うことができず，蓄積されたデータの解析を第三者に依頼して実施することも多いが，電子カルテデータのような個人を特定可能な機微な情報については，法令や契約によって，氏名や生年月日などの個人を特定可能な属性情報に対する匿名化が義務付けられている[7]．データを解析する際にはこれらの匿名化された属性情報を，個人を識別するための識別子として使用することが多いものの，これらの匿名化のため暗号化された属性情報は結婚や転職，退職などを契機として変化することがあり，即ち，識別子も変化することから，個人の追跡ができなくなる可能性がある．健康・医療分野に於けるビッグデータ解析では，健康状態からの疾患発症の予防や，薬剤の服薬に伴う副作用の検出，発症した疾患の増悪の予兆検出などの，長期間の時間軸のイベントについての解析への期待が高まっており[8, 9]，匿名化されたデータに於いても高い個人の追跡精度が求められている．

著者らは，我が国最大規模の健康医療データベースである特定レセプト情報・特定健診等情報（通称ナショナルデータベース(注1)）のデータのうち， 6 年分の電子レセプト情報について厚生労働省から提供を受け，解析サービスを医療分野の研究者に対して提供してきた[3]．本論文では，同省から提供を受けた 6 年分の電子レセプト情報に含まれる 2 種類の個人識別子を用いた個人追跡精度の検証を行うとともに，2 種類の暗号化された個人識別子に基づく個人の追跡手法を提案し，その識別可能性の検証結果を示した後，将来に向けた課題と展望を纏める．

本論文の構成は以下の通りである．2.では，電子レセプトと，電子レセプト情報及び特定レセプト情報・特定健診等情報データベースの概要，当該データベースに含まれる個人識別子の概要，個人追跡に於ける課題を述べる．3.では当該データベースに含まれる 2 種類の個人識別子に基づく個人追跡手法を提案し，4.にて提案手法の検証結果を示す．5.に於いて将来に向けた課題と展望を示した後，6.に於いて本論文を纏める．

2. 電子レセプト情報と特定レセプト情報・特定健診等情報データベース

2.1. 電子レセプト情報

我が国に於おける公的医療保険が適用される保険診療に於いて，保険医療機関（病院，診療所等，歯科医院等）及び保険薬局は，毎月，保険者に対する医療費の請求書を決められた様式で作成し，請求内容の適切性を判断する審査支払機関へ送付する．その後，審査支払機関での審査を行った上で，保険者への請求を行

(注1):国家レベルの悉皆性を備えたデータベースを意味して，ナショナルデータベース（national database; NDB）と通称されることが多い

い，保険者は医療費のうち保険者負担分を医療機関に支払う．この際の請求書は，レセプト（診療報酬明細書; medical insurance claim; （独）Rezept）と呼ばれ，請求元である医療機関の情報に加えて,被保険者毎に，患者の個人情報，傷病名，提供した医療行為，診療報酬ならびに参考情報等が纏められて，記載される．従来は，紙媒体を提出することで請求書処理が行われていたが,平成 11 年に磁気媒体等による電子的な手段による提出が認められ，審査支払機関でのレセプト電算処理システムの構築に伴い，電子化したレセプト情報を電子媒体で提出することが可能となった．その後，平成 27 年 4 月診療分からは原則，電子レセプトによる請求が原則化されており，平成 29 年 10 月期では，電子化率は医療機関数ベースで 93.3%，請求件数ベースで 98.2%に達している[1]．この際，レセプトは，任意の医療機関と任意の保険者の間で交換可能であることが不可欠であることから，電子化に際して，従来の紙媒体の様式を基に，電子的な様式の仕様が定められた．その後，政策や規則の変化に伴い，レセプトの様式仕様は改訂が繰り替えされている[6].

レセプトによる医療費の請求では，医療行為の対価である診療報酬は「出来高支払い方式」を基本とし，医療行為ごとに対価が定められており，医療機関が実際に行った医療行為の項目と数量に応じた医療費をレセプトに記載し，請求行う．ただし，一部の急性期病院への入院については，診断群分類に基づく定額報酬算定制度（diagnosis procedure combination / per-Diem Payment System, DPC/PDPS）が導入されており，当該制度に該当するレセプトは，医療行為ではなく臨床疾患エピソードによって診療報酬が決定される[12]．医療機関の種別等に応じて,記載様式が定められており，レセプトの様式は主に以下の 4 つに分類することができる．

- 医科用レセプト：病院や診療所等の医療機関に於いて患者が外来診療もしくは入院診療を受けた際に発行される．
- DPC 用レセプト：一部の急性期病院に於いて患者が DPC/PDPS 制度が適用された入院診療を受けた際に発行される．
- 歯科用レセプト：病院や診療所等の医療機関に於いて患者が歯科診療を受けた際に発行される．
- 調剤レセプト：調剤薬局に於いて患者が調剤を受けた際に発生する．

2.2. レセプト情報･特定健診等情報データベース

厚生労働省は，平成 20 年に改正された「高齢者の医療の確保に関する法律」（昭和 57 年法律第 80 号）に基づき，平成 21 年より今日まで電子レセプト情報及び特定健診・特定保健指導情報の継続的な収集を実施しており，収集した情報に基づき「レセプト情報・特定健診等情報データベース（通称ナショナルデータベース，NDB)）」[10, 11]の構築を行っている．平成 28 年度末では請求約 100 億件，レコード約 2,600 億件に達している．

レセプトは医療費の請求を元来の目的としているため，臨床検査に関する情報や診断画像等の臨床情報は含んでいない．しかしながら，我が国は国民皆保険制度を採用しているため，労災や自費治療等の例外的なケースを除き，国内で提供される医療行為の殆どは公的保険の対象となっている．そのため，国内のレセプトを一元管理する NDB は，国家レベルの悉皆性を有する我が国最大規模の診療データベースであるといえる．

2.3. 第三者提供用のレセプト情報･特定健診等情報データベースに含まれる個人識別子

厚生労働省は平成 23 年に「レセプト情報等の第三者提供に関する有識者会議」を設置し，データ利用に向けた「レセプト情報・特定健診等情報の提供に関するガイドライン」の整備を行う供に，平成 23 年度から試行的なレセプトデータの第三者提供を開始し，平成 25 年度から本格実施をしている．第三者提供の際には，レセプトに含まれる氏名や被保険者証の記号・番号，生年月日の「日」といった個人が識別可能な情報は削除され，代わりに，氏名や性別，被保険者証の記号・番号，生年月日を組み合わせた値に対してハッシュ関数を適用することで導出したハッシュ値を個人識別子として使用している[2]．第三者提供用の NDB に於ける個人識別子は以下の 2 種類である．

- ID1 : 保険者番号，被保険証の記号及び番号，生年月日，性別を使用し作成
- ID2：氏名，生年月日，性別を使用し作成

図 1 に個人に紐づくレセプトと各レセプトに記載される情報の例を示す．第三者提供用のレセプトには，個人識別子である ID1 と ID2，時間情報として医療提供した日付やレセプトを発行した年月等が記載される．

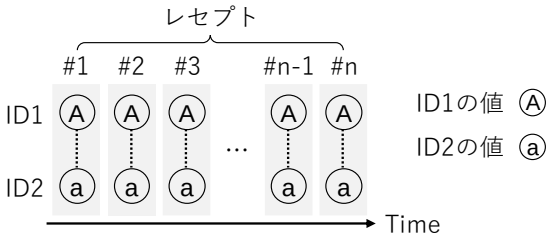

図 1　個人に紐づくレセプトと各レセプトに記載される ID1 と ID2 の例

図 1 に於いて，A は ID1 の値，a は ID2 の値の例をそれぞれ示しており，各レセプトには ID1 の値 A と ID2 の値 a の両方が記載される．また，各レセプトに記載された時間情報を使用することで，個々人に紐づく n 個のレセプトを#1, #2, #3, …, #n のように時間軸上に並べることができる．

第三者提供用の NDB を用いて，例えば，糖尿病患者に於ける糖尿病発症後の医療費の集計等の個人単位での医療エピソードを解析する際には，ID1 もしくは ID2 を個人識別子として使用し，使用した ID に紐づくレセプトを抽出することで解析が行われている．

2.4. 個人追跡に於ける課題

先述のとおり，第三者提供用の NDB に於ける個人識別子である ID1, ID2 は，保険者番号，被保険証の記号及び番号，生年月日，性別，氏名に基づき作成されている．そのため，ID1，ID2 は主に以下の理由によって値が変化し得る．

- ID1：転職，退職，扶養の適用などに起因する保険者の変更
- ID2：結婚などに起因する氏名の変更

また，医療機関に於ける患者情報の登録は職員による手入力で行われることが多いため，保険者情報や氏名の誤入力によって ID の値が変化する可能性がある．

図 2 に ID の値が変化した個人のレセプトの例を示す．ID1 もしくは ID2 のどちらか一方のみが変化した個人の場合には，変化していない ID を使用することで個人に紐づく全てのレセプトが抽出可能である．しかしながら，どちらの ID が変化するかは個々人によって異なり，また個人によっては，転職による保険者情報の変更が起きたのち，結婚等に伴う姓の変更が起こる，即ち，ID1 と ID2 の両方が変化する場合も存在する．特定の個人を追跡する期間が長くなるに従って転職や結婚などのイベントが発生する可能性は高くなるため，データの年度が増加するに従って ID1, ID2 が変化する可能性も高まり，個人に紐づく全レセプトを追跡できる可能性は低くなる．国家の医療政策の立案に繋がるデータ分析では，月単位の統計的集計などの短期的な時間軸の分析だけでなく，コホート調査等の数年単位のレセプトを使用した，長期的な時間軸の分析の両方が必要である．特に，疾患の増悪・治癒・寛解に伴う診療行為の時間変化などの長期的な時間軸の分析は，今後の我が国の医療資源の分配に於いて有益であると考えられるため，長期的な時間軸で個人のレセプトを追跡できることは重要である．そのため，今後もデータ量の増加が見込まれ，より長時間のレセプト情報が解析可能となる NDB を有効に活用するためには，ID 値の変化による個人追跡性への影響を抑制し，

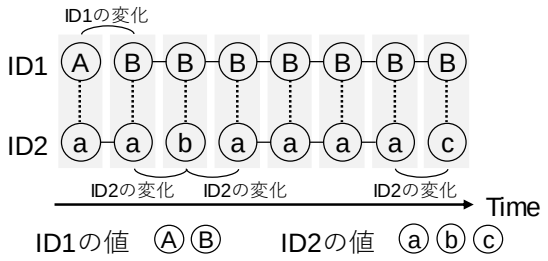

図 2　ID の値が変化した個人のレセプトの例

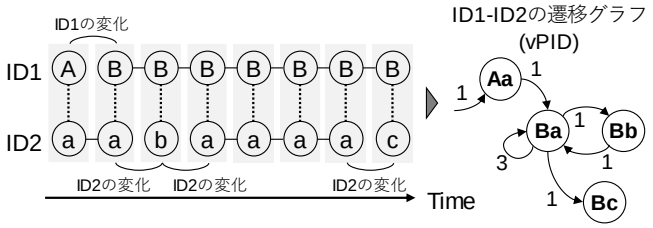

図 3　ID1-ID2 の遷移グラフ(vPID)の作成方法

長期間の時間軸での個人追跡を可能とする手法が必要である．

3. 2 種類の暗号化された個人識別子に基づく個人の追跡手法の提案

本論文では，2 種類の個人識別子に基づく個人追跡手法を提案する．本章では，提案手法である ID1-ID2 の遷移グラフを使用した個人追跡について述べたのち，作成した遷移グラフの特性の計測について述べる．

a)　ID1-ID2 の遷移グラフを使用した個人追跡

前述の通り，ID1 または ID2 のどちらか一方のみを使用して個人に紐づくレセプトを抽出した場合，転職や結婚等のライフイベントを契機として属性情報が変化した個人の全レセプトを追跡することができない．そこで本手法では，一度でも隣接する ID1 と ID2 の組み合わせを網羅的に抽出し，抽出した ID1 と ID2 の組み合わせの時間的な変化に基づき ID1-ID2 の遷移グラフを作成する．作成した遷移グラフに新たな識別子を付与し，この付与した識別子を仮想的な個人識別子（virtual patient ID, 以下 vPID）として使用することで，ID が変化した個人についても追跡を可能とする．図 3 に ID1-ID2 の遷移グラフの作成方法を示す．まず，一度でも隣接する ID1 と ID2 の組み合わせを網羅的に抽出し，レセプトの発行された日付に基づき並び替えを行う．その後，ID1 と ID2 の組み合わせを一つのノードとし，同一の ID の組み合わせが連続して発生している場合には自ノードへの遷移（以下，自己遷移），異なる ID の組み合わせが発生する場合には別ノードへの遷移として，それぞれの遷移頻度を集計することで遷移グラフの作成を行う．このように，レセプトに記

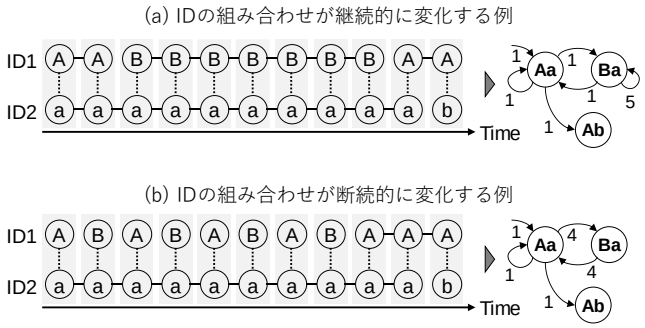

図 4　循環を有する遷移グラフの例

載されている 2 つの ID の両方を使用することで，ライフイベントを契機とした ID1 または ID2 の変化が生じた個人に於いても，変化していない ID を使用して個人を追跡することができる．

b)　ノード集約方式を使用した ID1-ID2 の遷移グラフの特性計測

通常，一度変化した個人の属性情報が変化前と同じ値となる，即ち，ID1-ID2 の遷移グラフ内に遷移の循環が発生する可能性は低いと考えられるが，以下の理由によって遷移の循環が発生するる．

- ID1：季節に応じた職業の変更，扶養の適用／不適用の繰り返し，遷移グラフ内に同姓同名・同生年月日・同性別の個人を複数含む
- ID2：離婚や死別などによる旧姓への変更，氏名の誤登録，遷移グラフ内に被保険者が同一の同姓の双子を含む

季節に応じた職業の変更や，扶養の適用／不適用の繰り返し，離婚や死別などによる旧姓への変更などが発生する確率と比較して，同姓同名・同生年月日・同性別の個人が存在する確率や被保険者が同一の同姓の双子が含まれる可能性は低いと考えられる．しかしながら，循環を含む遷移グラフの割合やその特性は未知である．そこで，提案方式で作成した遷移グラフに 2 種類のノード集約手法を適用することで，遷移グラフに含まれる循環の特性を計測する．

b-1) 方式 1 ：ID 遷移の頻度を活用したノード集約

図 4 に循環を有する遷移グラフの例を示す．一般的に，短期間に転職や氏名の変更が繰り返し発生する可能性は低い．そのため，同一の ID1 と ID2 の組み合わせを有するレセプトが連続して存在する，即ち，ID の組み合わせは継続的に変化すると考えらえる．（図 4(a)）．一方で，遷移グラフ内に複数の個人が含まれている場合，個々人がそれぞれのタイミングで医療を提供される．このような場合，異なる ID1 と ID2 の組み合わせを有するレセプトが交互に存在する，即ち，ID の組み合わせは断続的に変化すると考えらえる（図 4(b)）．そこで本方式では，ID の遷移頻度を活用することで，転職や氏名の変更に起因すると考えられる循環

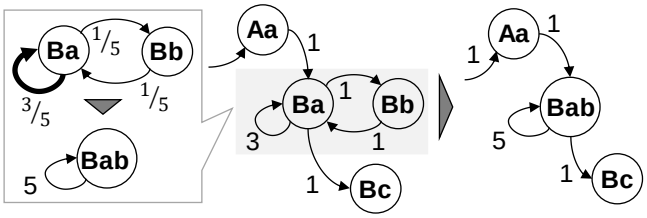

図 5　ID の遷移頻度を活用したノード集約方式の模式図

ノード集約方式1:
入力:
閾値：$threshold$
サイクルを形成するノード：$V = \{v_1, v_2, \cdots\}$
サイクル内の遷移の出現頻度　：$F = \{f(v_i, v_j)\}$
出力：サイクルを一つのノードに集約するか否か
初期値: $aggregation\_flg = false$
for n in V **do**
　if $(f(n,n)/\sum F) > threshold$ **then**
　　$aggregation\_flg = true$
　　break
　end if
end for
return $aggregation\_flg$

図 6　ID 遷移の頻度を活用したノード集約方式の疑似コード

の集約を行う．図 5 に ID 遷移の頻度を活用したノード集約方式の模式図を示す．上述の通り，遷移グラフ内に単一の個人のみが含まれる場合には，短期間に個人属性情報が繰り返し同じ値に変化する可能性は低く一定期間は同じ ID の組み合わせが継続して発生すると考えられる．その結果，ノード間の遷移頻度は低くなり，自己遷移頻度が高くなると考えれる．そこで本方式では，循環内の遷移頻度の総数と各ノードの自己遷移頻度を活用したノード集約を行う．図 6 に ID 遷移の頻度を活用したノード集約方式の疑似コードを示す．本方式では，循環を形成するノードに於いて，循環内の遷移の総頻度に対して自己遷移の割合が任意の閾値を上回るノードが存在する場合には，当該循環を構成する全ノードを 1 つのノードに集約する．本方式に於ける閾値を 1 とした場合，全ての循環は集約されない，即ち，処理なしと同値となる．また，閾値を 0 とした場合，全ての循環は集約される．

このように循環内の自己遷移頻度を活用したノード集約を行う事で，単一の個人のみが含まれている可能性が高い循環を集約することができる．

b-2) 方式 2；疾患情報を活用したノード集約

図 7 に疾患情報を活用したノード集約方式の模式図を示す．慢性疾患や生活習慣病などの疾患を発症した個人では，例えば，外来診療に於いて当該疾患について継続的な診療・検査・治療を行っている限り，医科レセプトに記載される疾患情報（疾患名および当該疾患の診療開始日）は同一となる．そのため，図 7 に示

すように同一の疾患情報を有するノードは，同一の個人から生成されたノードである可能性が高い．そこで本方式では，疾患情報の一致頻度を活用したノード集約を行う．図 8 に疾患情報を活用したノード集約方式の疑似コードを示す．本方式ではまず，循環を構成する任意の 2 つのノードを選択し，選択したノード間に於いて同一の疾患情報が存在する件数を集計する．一致する件数が任意の閾値以上となった場合には，選択した 2 つのノード間に疾患情報に基づくエッジを張ることで疾患情報グラフを形成する．前述の処理を循環に含まれるノードの全組み合わせに対して実施した後，循環を構成する全ノードが形成した疾患情報グラフに全て含まれ，かつ，疾患情報グラフが連結グラフである場合には，当該循環を構成する全ノードを 1 つのノードに集約する.本方式に於いて閾値を 0 とした場合，全てのノードが何かしらの疾患情報を有している場合にノードが集約される．

このように循環を構成する各ノードの疾患情報を比較することで，同一の個人から形成されている可能性の高い循環を集約することができる．

これらの 2 つのノード集約方式に於ける閾値を変化させることで，作成した遷移グラフに含まれる循環の特性を計測することが可能である．

4. 実験

4.1. 方法

本論文では，まず，NDB のレセプトデータから ID1-ID2 の遷移グラフを構築し，遷移グラフに対して新たな識別子を付与することで仮想的な個人識別子（virtual patient ID: vPID）を作成する．作成した vPID に於いて,循環を保有する vPID の割合を計測した後，2 種類のノード集約方式を使用することで，作成した vPID に含まれる循環の特性を計測する．

著者らは，平成 27 年に当該データベースから平成 21~26 年度の 6 年分の電子レセプト情報の提供を受けた．提供を受けるに当たり，被保険者の特定をより困難化するために匿名化レベルを上げる等の幾つかの加

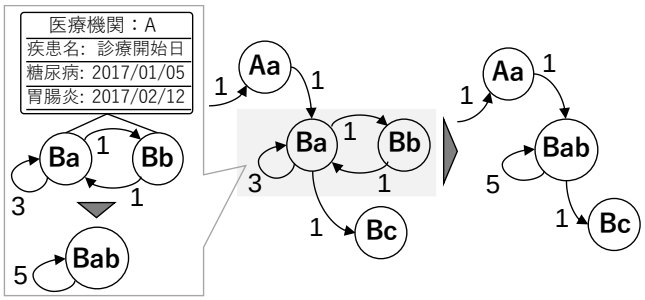

図 7　疾患情報を活用したノード集約方式の模式図

ノード集約手法 2:
入力:
閾値： $threshold$
サイクルを形成するノード：$V = \{v_1, v_2, \cdots\}$
単一の疾患情報： $d = \{$(疾患名：診療開始日)$\}$
ノードvの疾患情報：$D(v) = \{d_1, d_2, \cdots\}$
出力：サイクル内の全ノードを集約するか否か
初期値:
疾患グラフ： $G_d \coloneqq (g, V_d, E_d)$
疾患グラフのエッジの初期化： $E_d = \phi$
疾患グラフのノード：$V_d = \phi$
$aggregation\_flg = false$
for n in V **do**
  for m in V **do**
    if $n == m$ **then**
      continue
  $matched\_num == 0$
  for i in $D(n)$ **do**
    for j in $D(m)$ **do**
      if $D(n)(i) == D(m)(j)$ **then**
        $matched\_num++$
    end for
  end for
  if $matched\_num > threshold$ **then**
    E_dに (n, m) を追加
    V_dに n と m を追加
    E_dとV_dをもとにG_dを更新
  end for
end for
if $(V == V_d)$ && (G_dが連結グラフである) **then**
  $aggregation\_flg = true$
return $aggregation\_flg$

図 8　疾患情報を活用したノード集約方式の疑似コード

表 1　提供を受けた電子レセプト情報の諸元

レセプト種別	平成21年	平成22年	平成23年	平成24年	平成25年	平成26年
医科	654,668,008 件	884,069,925 件	921,452,091 件	943,053,734 件	955,043,583 件	972,275,439 件
ＤＰＣ	9,153,568 件	10,306,579 件	10,795,750 件	11,281,780 件	11,351,414 件	11,129,449 件
歯科	1,864,835 件	31,248,001 件	79,732,061 件	100,639,493 件	123,324,269 件	167,784,508 件
調剤	521,810,069 件	547,939,129 件	564,978,686 件	582,116,923 件	592,082,064 件	608,125,817 件
合計	1,187,496,480 件	1,473,563,634 件	1,576,958,588 件	1,637,091,930 件	1,681,801,330 件	1,759,315,213 件

工が行われたものの，6 年分の全ての電子レセプト情報が提供され，潜在的に，医療行為が提供された個人を網羅的に解析する機会を得ている．表 1 に提供を受けた電子レセプト情報の諸元を示す．本論文では，著者らが使用可能な 6 年分の医科，DPC，歯科，調剤レセプトを使用する．

4.2. 結果

表 2 に ID1, ID2, vPID の総数を示す．NDB に予め格納されている ID1, ID2 は，6 年間分のレセプトデータ中にそれぞれ約 2.73 億件，約 2.38 億件が存在した．一方で，6 年間分のレセプトデータから作成した vPID は約 1.38 億件となり，ID1, ID2 と比較して約 1 億件少ない件数となった.厚生労働省の平成 24 年の人口動態統計によると，平成 24 年度の人口は約 1.26 億人であり，各年度の出生数は 100 万人程度である[5]．人口と 6 年度分の出生数の合計値 1.32 億人と比較して，ID1, ID2 の件数は我が国の人口よりも数千万から一億件程度多い値となった．一方で vPID の件数は，人口と 6 年度分の出生数の合計値と近い値となったことから，vPID では保険者の変更や氏名の変更に伴う ID の変化を吸収できていると考えられる．

次に，循環を保有する vPID の割合を計測した．表 3 に 6 年分のレセプトデータから作成した vPID の内訳を示す．計測の際には，vPID に含まれる ID1 と ID2 が単一であるか複数であるか，vPID に循環が含まれる場合，当該循環が ID1, ID2 のいずれか若しくは両方の変化に起因する循環であるか，を軸として vPID を分類した．表 3 に示す通り，循環を含まない vPID は，［ID1-単一：ID2-単一］の約 3,468 万件，［ID1-複数：ID2-単一］の約 2,373 万件，［ID1-単一：ID2-複数］の約 306 万件，［ID1-複数：ID2-複数］の約 200 万件となり，その合計は約 6,374 万件と総 vPID 数の約 46%となった（表 3 太枠内）．その他の循環を含む vPID の合計数は約 7,477 万件と総 vPID 数の約 54%となった．

次に，2 種類のノード集約を使用した，vPID に含まれる循環の特性の計測を行った．図 9 にノード集約方式 1 に於ける閾値の変化に伴う循環を含まない vPID 割合の推移を示す．ノード集約方式 1 の閾値を 1.00 から 0 の間を 0.02 刻みで変化させた．その結果，閾値が約 0.20 となる時に，95%以上の vPID が循環を含まなくなった．図 10 にノード集約方式 2 に於ける閾値の変化に伴う循環を含まない vPID の割合の推移を示す．図 10 では，処理を行わない場合の値 46%を初期値とし，ノード集約方式 2 の閾値を 25（即ち，25 個の疾患情報が一致する場合にノードを集約する）から 0（何かしらの疾患情報を有している場合にノードを集約する）の間で変化させた．その結果，閾値を 10 まで変化

表 2　6 年間のレセプトデータに含まれる ID1, ID2, vPID の総数

識別子の種別	総数
ＩＤ１	273, 314, 230 件
ＩＤ２	228, 040, 615 件
ｖＰＩＤ	138, 245, 370 件

表 3　6 年分のレセプトデータから作成した vPID の内訳

6年分のレセプトデータから作成したｖＰＩＤの内訳			ＩＤ１		
			単一	複数	
			－	循環なし	循環あり
ＩＤ２	単一	－	34, 677, 448 件	23, 734, 828 件	6, 067, 476 件
	複数	循環なし	3, 059, 377 件	2, 003, 176 件	445, 984 件
		循環あり	27, 322, 463 件	28, 960, 718 件	11, 973, 900 件

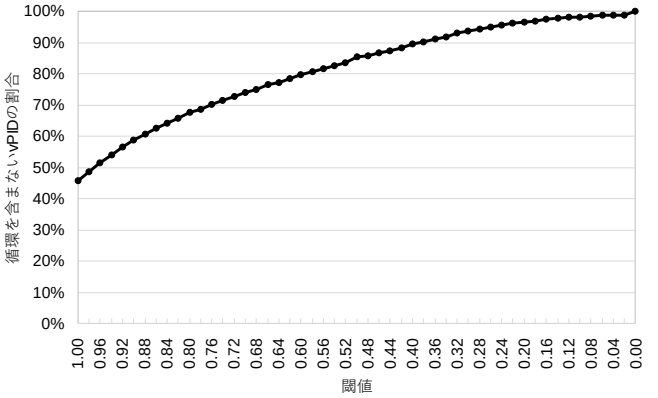

図 9　ノード集約方式 1 に於ける閾値の変化に伴う循環を含まない vPID 割合の推移

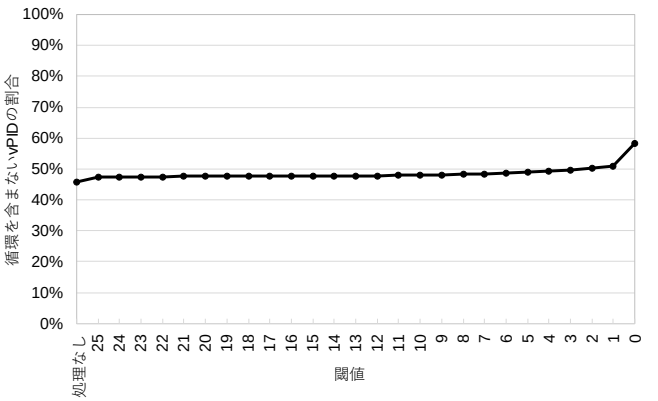

図 10　ノード集約方式 2 に於ける閾値の変化に伴う循環を含まない vPID 割合の推移

させたとしても循環を含まない vPID の割合は約 48%と初期値から 2%程度しか増加せず，閾値が 2 の時に約 50%，閾値 0 の時に 58%となった．図 11 にノード集約方式ごとの循環を含まない vPID 割合の比較グラフを示す．本比較ではノード集約方式 1 の閾値は 95%以上の vPID が循環を含まなくなる 0.20，ノード集約方式 2 の閾値は約 50%の vPID が循環を含まなくなる

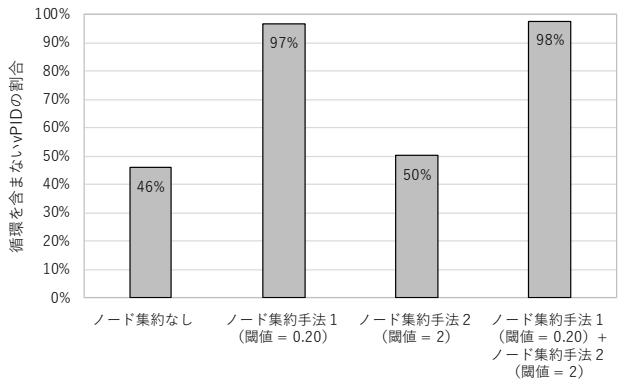

図 11　ノード集約方式ごとの循環を含まない vPID 割合の比較グラフ

2 とした．循環を含まない vPID の割合は，ノード集約を行わない場合で 46%, ノード集約方式 1(閾値=0.20)を適用した場合で 97%，ノード集約方式 2（閾値=2）を適用した場合で 50%, ノード集約方式 1(閾値=0.20)とノード集約方式 2（閾値=2）の両方を適用した場合で 98%となった．

ノード集約方式を用いない場合には循環を含まない vPID は全 vPID の約 46%であったが，ノード集約方式 1 と方式 2 の両方式を適応することで，循環を含まない vPID は約 98%まで増加することが判った．方式 1 に於いて閾値を 1.00 から 0.90 に変化させることで循環を含まない vPID の割合が 10%以上増加することから，数回のみ異なる ID に変化することに起因した循環が多く含まれていることが判った．このような突発的に別の ID に値が変化するような循環に対してノード集約方式 1 が有効であることを確認した．また，ノード集約方式 2 では，閾値が 0 の場合でも循環を含まない vPID の割合は約 58%に留まった．このことから残りの約 42%の vPID に於ける循環は，疾患情報が存在しないノードを含んでいることが判った．しかしながら，図 11 に示す通り，ノード集約方式 1 と方式 2 の両方を適用することで，方式 1 のみを適用した場合よりも循環を含まない vPID の割合が 1%(約 140 万件)増加することから，方式 2 についてもノード集約が有効であると考えらえる．

以上より，本提案により作成した vPID を使用することで, NDB に予め存在する ID1, ID2 よりも長期的な時間軸で個人のレセプトを追跡できることが示された．また，ノード集約方式を活用することで，vPID に含まれる循環の特性を計測できることが示された．

5. 課題と展望

本節では，今後の研究に向けた課題と展望を纏める．

a)　正解データセットを用いた提案方式の個人追跡精度の評価

本論文では，レセプト情報・特定健診等情報データベースから提供された 6 年分（平成 21~26 年分）の電子レセプトを使用し，個人追跡精度についての検証を行った．前述のとおり，提供された電子レセプトは個人情報保護の観点から，個人を特定可能な情報には暗号化が施されている．暗号化された個人識別子は個人の属性情報に基づき変化する可能性があるため，提供データには，個人を一意に追跡可能な識別子は含まれていない．そのため，本論文の提案手法を用いて作成した仮想識別子（vPID）を使用することで，予め含まれている ID1, ID2 と比較して，長期間の時間軸で個人のレセプトを追跡可能なことを示した. しかしながら，同一の vPID の中に含まれる個人が一意となるノード集約方式の閾値や，転職と同時に氏名が変更するなどの ID1 と ID2 の両方が同時に変化することにより追跡ができなくなる個人がどの程度含まれているか等の副作用については検証できていない．今後，個人を一意に追跡可能なデータを使用することで，本提案方式の詳細評価を行う事が課題である．

b)　実タスクへの適用

これまでに著者らは，医学系の研究者と共に，特定疾患の都道府県別患者数や，季節性疾患の月別発症率等の解析を行っている．これらの解析では，レセプト情報・特定健診等情報データベースに予め存在する ID1 または ID2 を使用して患者数の集計を行ってきた．しかしながら，これまで述べた通り，ID1 と ID2 は属性情報の変化を契機として ID の値が変化するため，実際の患者数よりも大きな ID 数が集計されてしまう，長期的な時間軸で患者を追跡できない，といった問題があった．前述の正解データを使用した本提案方式の詳細評価を行った後，我々が運用している解析システムに本提案方式により作成した仮想的な個人識別子（vPID）を組み込むことで，我が国の電子レセプト情報の分析精度向上に貢献し，社会的な便益に繋がる成果の創出に繋げていきたい．

6. おわりに

著者らは，厚生労働省が構築したレセプト情報・特定健診等情報データベース(ナショナルデータベース: NDB）から，平成 21 年度から平成 26 年度までの 6 年分に相当する匿名化された大規模電子レセプト情報の提供を受ける機会を得てから今日に至るまで，医学分野の研究者等に解析サービスを提供している．本論文では，レセプト情報･特定健診等情報データベース(ナショナルデータベース）に於ける個人追跡の課題を解決するため，2 種類の暗号化された個人識別子に基づく個人の追跡手法を提案し，検証結果について示すと共に，将来に向けた課題と展望を纏めた．

謝辞

本研究の一部は，厚生労働科学研究費政策科学推進研究「汎用性の高いレセプト基本データセット作成に関する研究」，厚生労働科学特別研究事業戦略研究「レセプト情報・特定健診等情報データベースを利用した医療需要の把握・整理・予測分析および超高速レセプトビッグデータ解析基盤の整備」，内閣府最先端研究開発支援プログラム（FIRST）「超巨大データベース時代に向けた最高速データベースエンジンの開発と当該エンジンを核とする戦略的社会サービスの実証・評価」，

内閣府革新的研究開発推進プログラム（ImPACT）「社会リスクを低減する超ビッグデータプラットフォーム」，日本医療研究開発機構（AMED）臨床研究等 ICT 基盤構築研究事業「エビデンスの飛躍的創出を可能とする超高速・超学際次世代 NDB データ研究基盤構築に関する研究」の助成に依る．レセプト情報・特定健診等情報データベースからの電子レセプト情報の第三者提供に掛かる手続きに関しては，厚生労働省保険局保険システム高度化推進室から丁寧なご指導を頂いた．

参　考　文　献

[1] 社会保険診療報酬支払基金．レセプト請求形態別の請求状況（平成 29 年度）: 平成 29 年 10 月診療分. http://www.ssk.or.jp/tokeijoho/tokeijoho_rezept/tokeijoho_04_h29.files/seikyu_2910.pdf. 2018 年 2 月 11 日に参照

[2] 厚生労働省．レセプト情報・特定健診等情報の提供に関するガイドライン．http://www.mhlw.go.jp/stf/shingi2/0000135204.html. 2018 年 2 月 11 日に参照

[3] 合田 和生，山田 浩之，喜連川 優，満武 巨裕．我が国の公的医療保険の悉皆分析を可能とする高速レセプト解析システムの開発と今後の展望．第 15 回日本データベース学会年次大会 (DEIM2017)

[4] 総務省．ビッグデータの流通量の推計及びビッグデータの活用実態に関する調査研究．http://www.soumu.go.jp/johotsusintokei/linkdata/h27_03_houkoku.pdf. 2018 年 2 月 11 日に参照

[5] 厚生労働省．平成 24 年（2012） 人口動態統計（確定数）の概況．http://www.mhlw.go.jp/toukei/saikin/hw/jinkou/kakutei12/dl/00_all.pdf. 2018 年 2 月 11 日に参照

[6] 厚生労働省保険局．診療報酬情報提供サービス．http://www.iryohoken.go.jp/shinryohoshu/. 2018 年 2 月 11 日に参照

[7] 厚生労働省，厚生労働分野における個人情報の適切な取扱いのためのガイドライン等．医療・介護関係事業者における個人情報の適切な取扱いのためのガイダンス， http://www.mhlw.go.jp/file/06-Seisakujouhou-12600000-Seisakutoukatsukan/0000194232.pdf

[8] 国立研究開発法人 科学技術振興機構 研究開発戦略センター．（調査報告書）医療・介護データ活用のための情報科学と社会基盤．https://www.jst.go.jp/crds/report/report04/CRDS-FY2016-RR-03.html． 2018 年 2 月 11 日に参照

[9] Ishikawa KB. ”Medical Big Data for Research Use: Current Status and Related Issues”， Japan Med Assoc J. 2016 Sep 1;59(2-3):110-124. eCollection 2016 Sep.

[10] 藤森研司．レセプトデータベース（NDB）の現状とその活用に対する課題．医療と社会, Vol. 26, No. 1, pp. 15{24, 2016.

[11] 満武巨裕，日本のレセプト情報・特定検診等データベース（NDB）の有効活用，情報処理, Vol. 56, No. 2, pp. 140{144, 2015.

[12] 厚生労働省. DPC／PDPS 傷病名コーディングテキスト．http://www.mhlw.go.jp/file/06-Seisakujouhou-12400000-Hokenkyoku/0000044471.pdf. 2018 年 2 月 11 日に参照