Vol.2013-DBS-158 No.31
2013/11/26

病院間のプライバシ保護データ連携による地域医療連携体制の評価フレームワーク

宮崎 昇幸[1,a)]　吉川 正俊[1,b)]　粂 直人[1,c)]　黒田 知宏[2,d)]　吉原 博幸[1,e)]

概要：医療資源の効率的な配分と利用は，とりわけ医療における構造的な問題を抱える国々にとって重要な課題となっている．これを背景に，地域における包括的な医療提供体制としての地域医療を確立する取り組みが，多くの国で注目を集めている．そのような医療体制では，地域に位置する医療機関が個々の機能や専門性を活かしながら相互に連携し，地域住民に質の高い医療サービスを提供することが求められる．これまで，各機関の生産性向上のための施策として，診療記録を電子的に取り扱うシステムが広く利用されるようになってきた．しかし，その運用によって蓄積される膨大な医療データは，地域医療の現状を評価し，把握するために十分利用されているとは言い難い．これは，従来の医療データの利用形態において，詳細な診療記録を患者のプライバシを脅かさずに病院間で共有することが困難であったことが要因となっている．そこで本研究では，地域医療の評価に注目した医療データの分析フレームワークを提案する．提案手法は情報理論的安全なマルチパーティ計算をベースとしており，医療データを各保有機関に分散したまま利用可能とすることで，詳細な診療記録を分析することに対する患者のプライバシ懸念を低減できる．本稿では，実際の医療データを用いた提案手法の評価に向けた，現状の課題とそれに対する対処法についても検討する．

1. はじめに

医療における資源配分や社会的背景にまつわる諸問題（e.g. 国民の高齢化による慢性疾患患者の増加，医療資源の地理的偏在，医師不足）は，世界各国で医療費増大などの根本原因として認識され，抜本的施策が求められている．これを背景に，抱える医療問題は国家間で様々ながら，世界的に共通して地域包括ケアと呼ばれる医療体制の整備が推進されているのは，興味深いことである．地域包括ケア体制とは，地域に位置する医療サービス提供者が個々の専門性を活かしながら連携し，地域住民に合理化された質の高い医療を提供する先進的な地域医療体制である．

医療機関を跨いだ連携を円滑に行うには，診療記録をはじめとする医療情報の電子化が不可欠である．医療情報の電子化は，個々の医療機関単位での取り組みにはじまり，EHR（Electronic Health Record）に代表される地域もしくは国家的規模での情報共有基盤を整備するに至る．地域包括ケアを推進する多くの国においては，機関ごとの電子化は広く浸透していながらも，システム間の連携は医療データの標準化と並行して整備が進められている段階である．

これに加え，地域医療体制の整備を先導する地方自治体にとっては，地域の現状と問題の把握，数値目標の設定，および施策の進捗の評価を行うための客観的な指標が必要となる．こうした指標計算のためには，病院情報システムの運用において蓄積される種々の医療情報（電子カルテなど．以後，包括的に医療データと呼ぶ）が素材として有用と見込まれる．しかし，現在のところ，それらのデータを地方自治体が一般的に利用することは困難であり，結果として地域医療に着目した評価体制も整備が遅れている．

これには，地域医療のための医療データ利用に求められる二つの要件が影響している．患者のプライバシ保護は，機微な情報を多く含む医療データを利用する際に不可欠な要件である．しかし，従来の利用形態では，保有機関外の第三者を信用してデータを共有しており，強いプライバシ懸念がデータ利用に対する合意形成の障害となっていた．

さらに，地域包括ケア体制の整備に特有の要件として，質の高い（多くの項目が詳細に記載されている）医療データを利用することが必要である．これは，機関横断的な比較・評価を考慮する場合，来訪患者の年齢や重篤度の偏りといった機関差を正しく反映しなくてはならないためであ

[1] 京都大学大学院情報学研究科
[2] 京都大学医学部附属病院
[a)] tmiyazaki@db.soc.i.kyoto-u.ac.jp
[b)] yoshikawa@i.kyoto-u.ac.jp
[c)] kume@kuhp.kyoto-u.ac.jp
[d)] tomo@kuhp.kyoto-u.ac.jp
[e)] lob@kuhp.kyoto-u.ac.jp

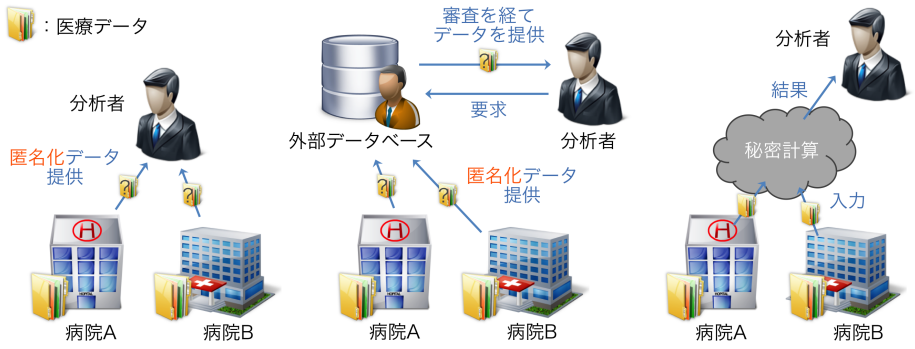

図 1 医療データの利用形態.

る．ところが，先述した従来のデータ利用形態では，匿名化 [*1] によるデータ改変や他の情報ソースとのリンケージ禁止制約が影響し，機関差を十分に精緻化可能な品質のデータを得ることが困難であった．

そこで本研究では，複数の機関に分散している医療データを他者に開示することなく利用することで，データの保護と高い品質を両立した分析を可能とする，地域医療体制の評価に適したフレームワークを提案する．ここでの「分析」とは，簡単な集計からデータマイニングに至るまで広範な計算を安全に行うことを指す．

本稿では，まず，医療連携の中で生成される種々の医療データの集合体が「地域医療連携の統合診療記録」を成すと捉え，地域医療の評価における分析対象としての概念と定義について述べる．次に，その部分データを各医療機関が分散管理する環境におけるプライバシ保護データ分析問題として，統合診療記録の分析を定式化する．続いて計算の手続きを述べ，最後に，入手可能な実医療データを用いた提案手法の評価のための課題や前処理について議論する．

2. 関連研究

2.1 医療データの利用形態

本節では，提案フレームワークの形態を含む三種類の医療データ利用アプローチ（図 1）を，パーソナルデータ保護とデータ品質の観点から概観する．

Model 1（分析者を信用）： データ保有者は，分析を行う第三者（分析者）を信用してデータを提供する．

Model 2（外部 DB を信用）： 外部のデータベースを信用して各保有者のデータを集約し，分析者は依頼・審査を経て提供された部分的データを利用する．

Model 3（保有者のみを信用）： データ保有者のみを信用し，他者にデータを開示することなく，協調計算によって得られる計算結果だけを分析者に提供する．

従来の医療データ利用形態は，Model 1 もしくは 2 に分類される．Model 1 では，保有者との間で秘密保持誓約を交わしたうえで，用途を限定する形で分析者にデータが提供される．この形態では，患者のプライバシ懸念が高まることが明白であり，かつ提供元の機関数も限られることが多い．そのため，広域的な視点での分析には不向きである．一方，Model 2 は，各国で整備が進められる地域または国家規模の医療情報データベースのモデルに該当し，広域かつ多数の機関から収集したデータを入手できるため，これまで医療政策などの研究に多く利用されている [7]．

ただし，これらのモデルで提供されるデータは，個人識別可能性のある属性の匿名化（年齢を 5 歳刻みで記録するなど）や稀少データの排除など，レコードごとの特異性を平滑化する改変が施される．さらに，このような修整に加え，入手データと他の情報ソースとのリンケージを禁止する制約が設けられる場合が多い．したがって，入手データに記載されていない項目は利用できず，先述の改変とあわせてデータの品質を落としてしまう．このようなデータ品質上の制約は，特にデータのプライバシ保護と高い品質を両立することが要求される分析を困難とし，我々が注目している地域医療体制の評価に適用することは難しい．

本研究のアプローチは，データを保有する医療機関にのみを信用する Model 3 に該当する．医療機関は暗号技術をベースとする秘匿協調計算を実行し，得られた結果だけを分析者に提供する．一般に，秘匿計算技術は多大な計算コストを要し大規模データ上の分析には適さないが，プライバシ保護とデータ品質の両立が可能である．このことから，本アプローチによって，地域包括ケアの整備をはじめとする医学領域の発展のための新たな医療データ利活用の方向性を示すことが期待される．

2.2 医療体制改善のための医療データ分析

医療の質の向上に有用な知識を発見することを目的として，医療データのマイニングを行う研究は盛んに行われてきた．Alshwaier と Emam[1] は，複数の医療機関に分散した医療データを分析利用することに着目し，種々の

[*1] データの個人識別性を低減するため，属性の削除や数値の丸めを行う統計的プライバシ保護処理.

匿名化手法の比較と最適なアルゴリズムの検討を行っている．Zaidi[21] は，互いに形式の統一されていない複数の医療データリポジトリにおける，エージェントベースの分散データマイニングフレームワークを提案している．この手法では，各リポジトリからデータを収集する Data Collection Agent が介在しており，データを分散したまま利用する本研究の目的とは異なっている．

また，医療政策に関する研究においても，電子レセプト *2 データをはじめとする医療データが盛んに利用されている．伏見と松田 [8] は，case-mix 分類 *3 を付加した患者調査データを用いて，地域における医療資源の必要量の推定や，患者の受療行動の視覚化を行っている．藤森 [7] は，北海道内で匿名化して収集された電子レセプトデータを分析し，患者の病態別受療動向や，地域における医療機関の連携を可視化することを試みている．これらの研究で用いられるデータは，匿名化と利用制約のもとに提供されたものである．しかし，地方自治体や医療機関による地域医療の PDCA サイクルを機能させるためには，より安全かつ一般的な医療データ利用基盤が必要となる [14]．

2.3 プライバシ保護データ分析

プライバシ保護データ分析（Privacy-preserving Data Analysis; PPDA）は，医療データの利活用において重要な要素技術である．PPDA によって，データのプライバシを脅かすことなく機微なデータの分析を実行できる．

PPDA は，データ変換と秘匿計算という二つのアプローチに分類することができる．データ変換アプローチでは，必要な統計的性質を残しながら個人識別に繋がる情報を排除し，データの開示によってプライバシが侵害されないことを保証する．識別不能なデータを必ず k 個以上含むことを保証する k– 匿名化 [18] は，最もよく知られた手法の一つである．ただし，データ改変によるプライバシ保護の度合いと分析結果の正確さの間にはトレードオフ関係が存在し，これらを両立するには事例ベースの議論が必要となる．

一方，秘匿計算アプローチは，暗号技術をベースとする．マルチパーティ計算（Secure Multi-party Computation; SMPC）[10][3] は，複数のパーティが持つ秘密情報を入力とする任意の関数計算を，第三者機関の補助なしに安全に実行するプロトコルである．SMPC は，計算の汎用性が高く厳密な結果を得ることができるが，計算コストが高いという難点がある．他には，準同型暗号を用いた実装も存在する．準同型暗号は，暗号文のまま演算ができる性質を持つ暗号であるが，これまで任意回の加算と乗算を効率的に実行可能な手法は知られておらず，汎用性を犠牲に特定の計算に特化したプロトコルを構成するために用いられる．

本研究では，SMPC に基づくアプローチにより，データを分散したまま種々の分析を安全に行う．目的の類似した研究としては，Drosatos と Efraimidis[6] が，健康器具から収集されるセンサ情報の分散 PPDM を行う加法準同型暗号ベースのマルチエージェントシステムを提案している．

3. 地域医療体制の評価

3.1 評価の目標

これまで，地域における医療連携体制を評価する取り組みはあまり進んでいない．紹介率/逆紹介率や再発率，死亡率，平均在院日数といった評価指標も検討されてきたが，それらは医療機関に対して体制改善への動機付けを与えるには至っていないのが現状である．これは，複数の医療機関から詳細な医療データを収集することが困難であり，各機関の専門性や来訪患者の性質といった差異が適切に評価に反映されていなかったことが原因となっている．

一方，地域医療の評価に向けた医療データの利用可能性と，候補指標を示している研究も存在する．図 2 は，藤森の研究 [7] において匿名化電子レセプトデータより抽出された指標を表し，ある地域における脳卒中医療の機関間連携とパス別の患者割合を可視化したものである．この結果から，複数の急性期病院から患者を受け入れる回復期病院の存在や，病院 E と E' の緊密な連携など，従来よりも詳細な地域の医療状況が観察できる．

ここで，先行研究で利用されてきた電子レセプトをはじめとする医療データは，病院情報システムに蓄積される豊富な診療記録の一部を目的に応じて切り出した部分的データとして捉えることができる．例えば，レセプトは医療費請求のための請求書であり，行われた検査の結果に関する情報は含まれない．これに対し，提案フレームワークでは，各機関のシステムが有する大元の診療記録データを利用し，データの内容に関する制約を低減することを目指す．

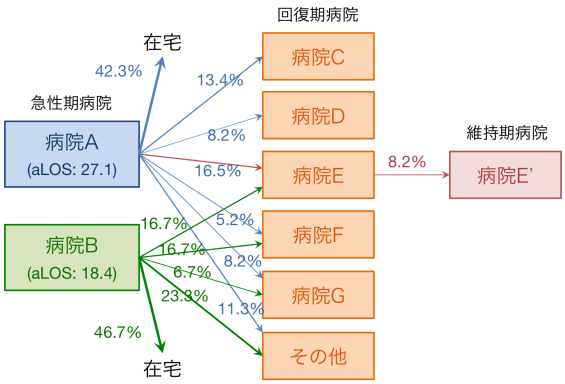

図 2 ある都市における脳梗塞（JCS < 30）の地域連携（文献 [7] より引用） – JCS（Japan Coma Scale）は意識障害の評価指標，aLOS（average Length of Stay）は平均在院日数を指す．

*2 レセプト（診療報酬明細書）とは，医療機関が行った診療行為について保険機関に請求する医療費の明細書をいう．電子レセプトとは，電子的に請求・審査を行うために電子データ化されたレセプトを指す．
*3 共通コードを用いて，重症度などに基づき患者を分類する方法．

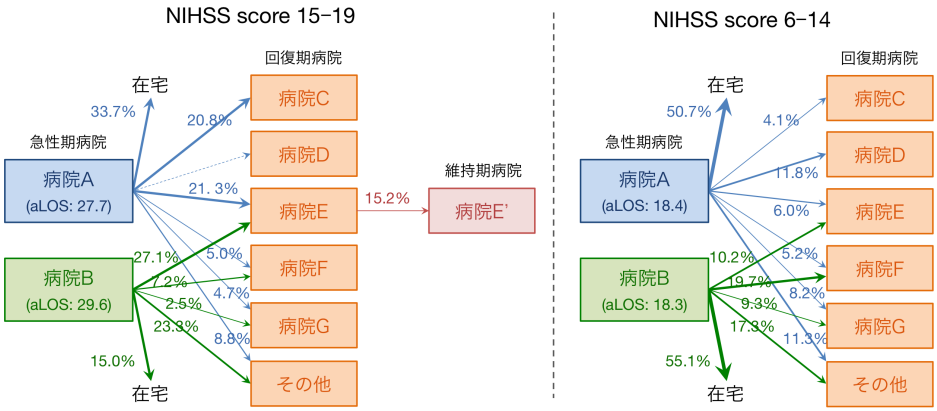

図 3 詳細化された地域医療の評価指標のイメージ図.

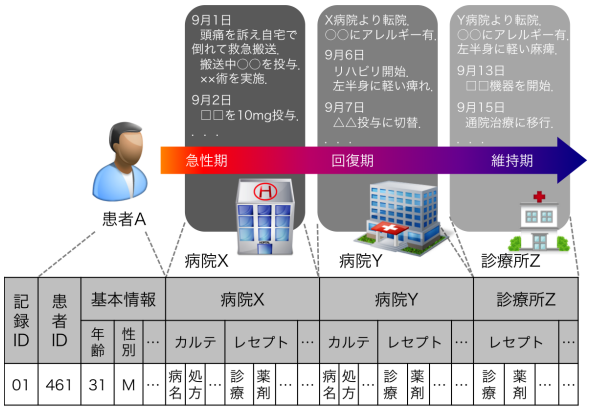

図 4 統合診療記録の概念図.

図 3 は，図 2 の指標を患者の NIHSS スコア [*4] によってさらに分類したイメージ図を表し，このように高品質なデータを利用した詳細な評価を本研究の目標とする．

3.2 分析対象

提案フレームワークにおける分析対象は，ある患者に対する連携医療の中で複数の医療機関で生じる種々の医療データである．本研究では，そのようなデータの集合が，ある患者の発症から回復，在宅維持に至るまでの連携医療の記録を成していると捉え，これを「地域医療連携の統合診療記録」(Integrated Medical Records of Regional Medical Cooperation; IMR) と呼ぶ．その概念図を，図 4 に示す．図中のテーブルは，ある患者の受療記録を表す IMR レコードに，患者に関する基本情報を付加したものである．

様々な形式の医療データを統合することで，IMR には相互に補完されたデータ項目が含まれ，豊富な情報を分析に利用することが可能となる．現実には，IMR はいくつかの部分記録に分解して複数の医療機関に分散され，様々な形式で保持されている．IMR のレコードは，ある患者に対する初診から在宅維持に至るまでの一連の医療行為を単位とし，一意な記録 ID によって識別される．IMR レコードに包含される医療データ群には，親レコードの記録 ID が付与されているとする．また，個々の患者と医療機関にも，全機関を通して一意な患者/機関 ID が付与されているとし，識別はこの ID によって行う．

各医療機関が保有する医療データは，機関ごとの傷病名などの表記ゆれや，分割して記載すべき項目が統合されている（低分解能と呼ぶ）場合があるなど，分析に利用しにくい性質を有している．また，各項目に記載されうる要素数は事例・個人によって様々であり，医療情報そのものの性質も複雑である．本研究では，記載内容の表記ゆれは考慮しない（荒牧ら [2] の手法などにより，事前に修整可能である）が，名前空間の相違や低分解能は想定する．したがって，これらのデータの表現には，柔軟な構造を持つデータ形式を利用することが望ましい．

そこで，提案フレームワークでは，これらの医療データを XML（Extensive Markup Language）形式で表現し，記録 ID や患者情報など固定的な項目をリレーショナル列，医療データを XML 列に格納するハイブリッド型 XML データベースで IMR レコードを取り扱う．ハイブリッド型 XMLDB は，関係データベース（Relational Database; RDB）に XML をそのまま格納して検索可能としたものであり，大規模データに対する高速な一括処理を利点とする RDB と，まばらな属性の取り扱いや柔軟な項目の追加/削除に長ける XML の長所を活かした構造である．

現状では，医療データのデータ形式も機関ごとに統一されておらず，共通の XML スキーマに格納する標準化処理が必要である．このような統一形式 XML としては，HL7 v3（Health Level Seven Version 3）XML[11] や MML（Medical Markup Language）[19] などの医療情報交換共通規格の利用が考えられるが，必要な分析に最適な形式を逐一選択することも考えられるため，ここでは単一の形式に限定しない．本稿では，医療データは各機関において RDB，もしくはハイブリッド型 XMLDB で管理されていることを想定して議論を進める．分析に利用するデータ形式の標準化については，第 5 章にて言及する．

[*4] 脳卒中に起因する損傷を定量的に評価するための評価スケール．0 から 42 のスコアで重篤度を表す．

4. マルチパーティ計算

統合診療記録の秘匿分析には，Shamir の (k,n) 閾値秘密分散法 [17] をベースとするマルチパーティ計算（SMPC）[3] を利用する．以下では，(k,n) 閾値法を紹介し，それに基づく加算と乗算の SMPC プロトコルを述べる．

4.1 (k,n) 閾値秘密分散法

秘密分散法は，情報に秘匿性と対災害性を持たせて分散管理する方法として，Blakley[4] と Shamir[17] によって独立に提案された．以下に，Shamir による (k,n) 閾値秘密分散法の手続きを示す．ここで，n 人の参加者 $P_1,\cdots,P_n$ はそれぞれ識別子 $1,\cdots,n$ を持ち，かつ $k \leq n$ である．

(1) 秘密 s を知るディーラ D は，大きな素数 q を $q > \max(s,n)$ となるように選ぶ．

(2) D は，$f(0)=s$ となるランダムな $k-1$ 次多項式 $f \in \mathbb{F}_q$ を選ぶ．

(3) D は，シェア $s_i = f(i) \mod q$ を各 i $(1 \leq i \leq n)$ に対して計算し，P_i に与える．

(4) 参加者 k 人分のペア (i,s_i) を集めることで，Lagrange 補間によって多項式 f を復元できる．さらに $f(0) \mod q$ を計算することにより，秘密 s を復元できる．

4.2 基本計算プロトコル

(k,n) 閾値法をベースとするマルチパーティ計算は，各参加者がシェア上の部分計算を個々に行うことで，データを秘匿したまま任意の計算の結果を得ることを可能とする．任意の関数計算は，加算と乗算を組み合わせた回路を構成することによって実現される．以下に，基本となる加算と乗算のプロトコルを示す．ここでは，先述の (k,n) 閾値法において，秘密 s_1,s_2 が $k-1$ 次多項式 f,g によって n 人の参加者に分散されているとする．各計算は素数 q を法として行われるが，簡略化のため “mod q” は省略する．

各参加者 P_i は，シェアの和 $f(i)+g(i)$ を個別に計算することで，秘密の和 s_1+s_2 に対応するシェアを容易に得ることができる．実際，多項式 f,g の定数項はそれぞれ s_1,s_2 であるから，その和 $f(x)+g(x)$ は定数項 s_1+s_2 を持ち，$f(i)+g(i)$ はそのシェアである．したがって，$(i,f(i)+g(i))$ を k 人分集めて復元できるのは，s_1+s_2 となる．

同様に，秘密の積 $s_1 * s_2$ に対応するシェアは，$f(i)*g(i)$ を個別に計算することで得られる．ただし，このとき多項式 $f*g$ は $2(k-1)$ 次であり，かつランダムでもない．以下に，$f*g$ の次数下げ，ランダム化を含んだ積の秘密計算プロトコルの一つである Gennaro ら [9] の手法を示す．ただし，ここでは $n \geq 2k-1$ が前提となっている．

(1) 各参加者 P_i は，$f(i)*g(i)$ を個別に計算する．

(2) P_i は，$h_i(0) = f(i)*g(i)$ となるランダムな $k-1$ 次多項式 $h_i \in \mathbb{F}_q$ を選び，シェア $h_i(j)$ を各 P_j に送る．

(3) P_i は，受け取った $(1,h_1(i)),\cdots,(n,h_n(i))$ の Lagrange 補間によってシェア $H(i)$ を計算する．ペア $(j,H(j))$ を k 人分集めて復元できるのは，$s_1 * s_2$ となる．

5. 統合診療記録のプライバシ保護分析

5.1 準備

提案フレームワークを構成する主体は，分析を依頼するクライアント C，データの所有者である医療機関 $M_1,\cdots,M_n$，および計算を主導する管理サーバ S である．このうち，$M_1,\cdots,M_n$ が，秘匿協調計算を実施するエージェントとして機能する．ただし，ベースとなる秘密分散法の閾値 k (≥ 2) に対して $n \geq 2k-1$ であり，S は，PPDA 問題における *trusted third party* ではなくクライアントとエージェントの仲介者として機能する．また，クライアント C は $M_1,\cdots,M_n$ の中に存在してもよい．

各エージェントは，semi-honest モデルに従って振舞うとする．semi-honest モデルでは，エージェントはプロトコルで定められた振る舞いから逸脱しないが，実行過程において自身が受け取る全メッセージを蓄積し，相手の情報を推測しようとする．「定められた振る舞いから逸脱しない」とは，他のエージェントとの結託やメッセージの改変といった，能動的な攻撃を行わないことを指す．全ての参加者間には安全な通信路が存在し，なりすましや盗聴が無いことを保証されているとする．

PPDA の研究では，入力データを図 5 のような三種類のデータの分割モデルに分類して議論される．垂直分割は，全データエントリにおける属性の部分集合が，分割されたデータセットを構成する．水平分割では，全属性におけるデータエントリの部分集合がデータセットを構成する．本研究が対象とする地域医療連携の統合診療記録は，両者が混在した任意分割モデルに該当する．

地域における医療連携は，図 2 に示すように，一般に n 対 n 関係となる．したがって，地域医療連携の統合診療記録（IMR）のレコードには，複数の医療機関が様々な順序

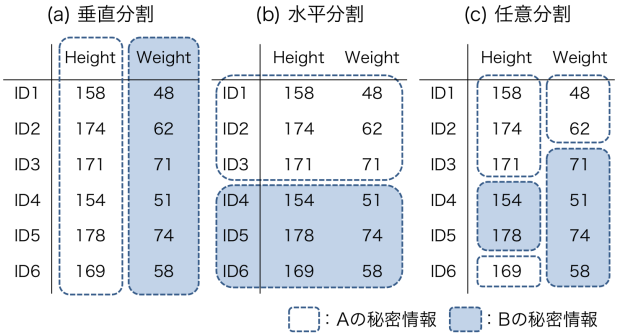

図 5 データ分割モデル（文献 [16] より引用）．

表 1 IMR テーブルの属性.

(a) 病期に基づく格納

記録ID	…	急性期	回復期		維持期
01	…	病院A (XML)	病院B (XML)	病院C (XML)	診療所D (XML)

(b) 診療順序に基づく格納

記録ID	…	病院1	病院2	病院3	病院4
01	…	病院A (XML)	病院B (XML)	病院C (XML)	診療所D (XML)

で関与した記録が混在することになる．IMR テーブルの属性として，急性期，回復期，維持期といった病期を付与する場合（表 1a），そこに格納される医療データ保有者の順序は様々であり，かつ単一の列に複数の機関が関与することも許容しなくてはならない．以後，IMR テーブルにおける医療データの格納属性は，受療順序が保存される形で表 1b のように設定するが，事例ごとに適した構成の検討は今後の課題とする．

5.2 分析手続き

IMR のプライバシ保護分析は，以下の手順からなる．

1. **依頼** クライアント C は，管理サーバ S に結果を得たい計算内容を送信し，分析を依頼する．S は依頼を受け取ると，機関 $M_1, \cdots, M_n$ に，医療データを格納すべき XML の統一スキーマと写像スキーム，および依頼内容をブロードキャストする．
2. **検索** M_i は, 依頼内容をもとに自身のデータベースを検索し，記録 ID をキーとする部分レコード $R_1, \cdots, R_m$ を得る．$R_1, \cdots, R_m$ の医療データ列は，S から受け取った写像スキームに基づき統一形式 XML に変換する．このとき，空値のインスタンスが存在してもよい．
3. **分散** M_i は，$R_1, \cdots, R_m$ 中の各インスタンス（XML 列）について n 個のシェアを生成し，レコード単位で $M_1, \cdots, M_n$ にシェアする．
4. **IMR テーブルの再構成** M_i は，受け取ったレコードを記録 ID に基づいて結合し，IMR テーブルを再構成して自身のハイブリッド型 XMLDB に格納する．
5. **計算** IMR テーブル上での秘匿分析を，$M_1, \cdots, M_n$ が協調して行う．
6. **集約** $M_1, \cdots, M_n$ の計算結果を S に集約し，結果を復元する．このとき，各機関が保有する IMR テーブルは破棄してよい．S は，得られた結果の開示可能性を検証し，問題が無ければ C に開示する．

5.3 XML への写像

ここでは，検索フェーズで行われる，医療データの統一形式 XML への写像について述べる．この写像により，医療データの表現における機関ごとのネーミングルールや名前空間，型・構造などのばらつきを標準化する．

本稿では，これらの医療データが，各機関において RDB もしくはハイブリッド型 XMLDB で管理されていることを想定しているため，関係表から XML，および XML から XML への写像について論じる．このほか，病院情報システムで利用されることがあるオブジェクト指向 DB などからの写像は，文献 [15] などを参照されたい．

RDB 上のデータを XML データに写像する手続きは，主に以下の 4 ステップからなる [20].

(1) 関係スキーマを非正規化し，結合された表に戻す.

(2) RDB のリレーションを，XML の木構造に写像する.

(3) 関係スキーマを，XML スキーマ（DTD）に写像する.

(4) RDB のデータを，XML に写像する.

XML 間の変換は，XSLT（Extensible Stylesheet Language Transformation）によって実現可能である．XSLT では，XML 文書であるスタイルシートにオプションを記載し，変換したい XML 文書とともに変換メソッドに入力することで，変換後の文書が出力される．

この際，統一スキーマにおける XML インスタンスの粒度に合わせて，写像元データ項目の分割もしくは統合が必要になる場合がある．そのためには，各機関が記録内容の特性に合わせた自然言語処理等を行う必要があり，事例ベースの議論を要する．

5.4 個人匿名性の保証

分析結果の表現方法としては，対象条件で分類された表形式や，統計情報化した数値が考えられる．特に前者のようなミクロデータとしての性質を残す場合には，少なからずプライバシの侵害に繋がるリスクを有することになる．先述の分析手続きでは，集約フェーズで復元するまで結果値は秘匿されており，計算フェーズの後処理として分析結果のプライバシを保証することができれば有用である．

ここでは一検討として，SMPC ベースの秘匿クラスタリング（関連：[12]）を利用した $k-$ 匿名化手続きを示す．

サイズ固定 k-means クラスタリング 以下の手順を繰り返し，最も類似した k データをクラスタに分類する．

(1) 参加者は, 既存クラスタ C_i の中心 μ_i を協調計算する.
(2) 参加者は, 各 μ_i と各データ x_j の距離を協調計算する.
(3) 参加者は，各データ x_j について，現在のクラスタと他の最善のクラスタまでの距離の差を協調計算する.
(4) 3 で計算した距離順にデータをソートし，改善の大きいものから再割り当てを行う．移動先クラスタの個数制約を破らない場合，そのまま割り当てる．個数制約が破られる場合，2 つのクラスタでデータを交換し，最適な割り当てに近づける.
(5) 局所最適な割り当てにたどり着いたら，終了.

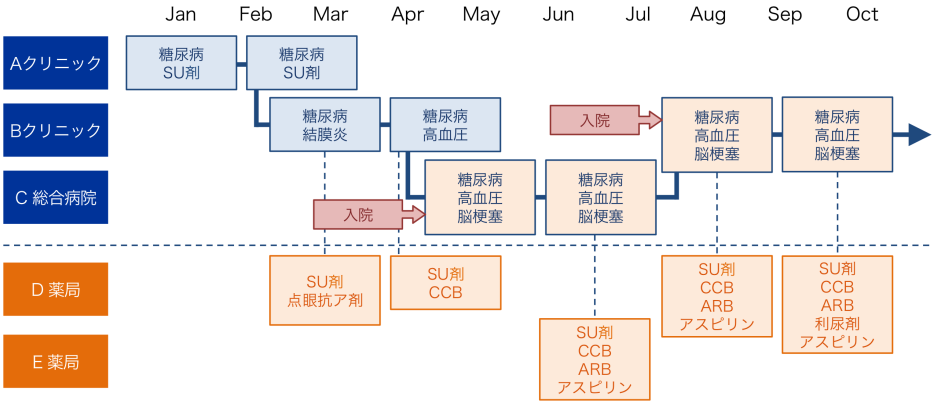

図 6 電子レセプトに記録された患者の受診履歴（文献 [13] より引用）.

Microaggregation[5] による匿名化 クラスタリングした k 個のデータを，協調計算した平均値などの代表値で置換することで，$k-$ 匿名性 [18] を保証する．

ただし，ここでは量的な属性を想定しており，質的な属性に適用するには何らかの距離算出手法が必要となる．

6. 性能評価環境の整備

本章では，実医療データを利用した提案手法の評価に向け，評価環境構築の課題とアプローチについて述べる．

以降では，例として Japan Medical Data Center (JMDC) Claims Data Base[*5][13] より提供される電子レセプトデータセットを用いる．同データから入手可能な情報の一例は，以下の通りである．

- 医療サービスの実施日，治療の頻度，期間．
- 患者の年齢，性別．
- 医療機関の病床数，経営形態，専門性．
- 病名．
- 医療行為の内容（検査，投薬，処置，リハビリなど）．
- 医療費．

電子レセプトデータは，請求元医療機関情報に基づいてデータが分散保持されている環境を再現でき，本研究のデータ保有モデルを擬似的に構築するには適している．さらに，医療機関に跨がった患者名寄せを施すことで，個々の患者の受診履歴を追跡することが可能である．JMDC Claims Data Base のレセプトデータは，あらかじめ患者名寄せ，匿名化，およびエントリの表記揺れの標準化を施したうえで格納されている．

提案フレームワークの理想的な運用環境としては，各医療機関が個々の患者に対する治療に関するあらゆる情報（e.g. 病気，病歴，紹介元/先の医療機関）を電子的に保持していることを想定している．しかし，現在我々にとって利用可能な実医療データは，主として第 2.1 節の Model 2 に該当する手段で入手するものであり，品質に制約が存在する（第 2.1 節および第 3.1 節を参照のこと）．特に，これらのデータには評価環境構築に必要となる医療連携の情報が記録されておらず，ある一連の医療行為が医療連携によるものか否かを抽出する必要がある．

第 3.1 節にて述べたように，データリンケージに関する制約から，入手データの品質を補うことは容易ではない．これに対し，評価指標の計算によって提案手法の有用性を示すためには，現実的な情報を人工的に付与し，高品質なデータを擬似的に作成することが考えられる．例えば，本来検査結果の情報を含んでいない電子レセプトデータの品質は，人工的な検査データを付加することで擬似的に向上させることができる．現実性については慎重な議論が必要であるが，そのような疑似データを利用することで，地域医療の評価に向けた展望を示すことが可能となる．

また，提案手法の評価環境は医療データの分散と医療連携に関する情報に大きく依存しているため，IMR の抽出は非常に重要な課題である．先述のように，患者の紹介/被紹介など医療連携に関する記録がない場合には，データセットに記録された一連の医療行為が連携医療によるものかを推定する必要がある．ここで，複数のレセプトデータからは，図 6 に示すように患者の受療行動を時系列的に追跡することができる．そのため，直感的な IMR の抽出手法としては，行為の時間的な連続性を利用することが考えられる．しかし，実際には，紹介状を受けた患者は即座に次の機関に受診するとは限らない（軽度の慢性疾患患者では，その間隔が数ヶ月にもなる場合がある）．したがって，IMR の抽出は，記録の時間的連続性と疾病情報に基づいて行う．一般に電子レセプトデータには，主傷病名，およびレセプト病名 [*6] と呼ばれるものなど複数の病名が含まれているが，ここでは記載の信頼性の観点から主傷病名を利用する．ここで，複数の機関における時間的に連続した同一の疾病に対する治療は，連携医療によって能動的に，もしくは患者の受療行動によって起こる可能性があるが，今

*5 株式会社日本医療データセンターが提供する医療情報データサービス．国内の健康保険組合から収集した医科（入院・入院外），および調剤レセプトを蓄積．

*6 保険請求上，治療行為や検査を正当化するため記載される病名．

回は両者を区別しない．

7. 検討と今後の課題

7.1 準同型暗号アプローチとの比較

本稿では，PPDA 手法として秘密分散法に基づくマルチパーティ計算を利用したが，準同型暗号によるアプローチも考えられる．準同型暗号を利用する場合，特定の演算に特化した高速な構成が可能だが，SMPC のような計算の汎用性は失われる．一般に，SMPC では高い計算コストが問題となるが，想定される評価分析は即応性を要求されるものではなく，計算の汎用性を優先すべきと考えられる．

7.2 事例ベースの具体化と性能評価

本研究の喫緊の課題は，提案フレームワークの性能評価である．第 6 章に基づいて実医療データを用いた評価環境を構築し，XML への写像，秘匿計算といった分析に係るコストの評価を行うことが必要である．

また，本稿では，XML の粒度にあわせた項目の分割/統合といった標準化手続きを具体化しておらず，検討が必要である．写像後の XML に存在しうる空値インスタンスについても，分析事例に応じて取り扱いを議論する必要がある．第 5.1 節で言及したように，IMR テーブルの格納属性についても，評価事例ベースの更なる検討が必要である．

これらの議論は，事例に注目して進めることが有効であると考えられるため，いくつかの分析事例を選定し，必要となる医療データ収集を緊急に進めることが必要である．

7.3 地域医療の評価指標

本研究の最終的な目的は，提案フレームワークが医療体制の整備や政策学的研究を進展させる中核技術となりうると示すことである．そのためには，今回行ったような性能評価だけでなく，実際に計算した地域医療の候補指標を医療分野に対して示すことが必要である．この研究では，候補指標から革新的かつ有用な指標が見いだされること，および医学分野からの更なる後押しを期待しながら，引き続き医療分野との連携によって指標選定・計算に取り組む．

参考文献

[1] Alshwaier, A. and Emam, A.: Data Privacy on E-Health Care System, *International Journal of Engineering, Business and Enterprise Applications*, Vol. 3, No. 2, pp. 89–99 (2013).

[2] Aramaki, E., Imai, T., Miyo, K. and Ohe, K.: Orthographic Disambiguation Incorporating Transliterated Probability, *International Joint Conference on Natural Language Processing 2008*, pp. 48–55 (2008).

[3] Benaloh, J. and Leichter, J.: Generalized Secret Sharing and Monotone Functions, *Advances in Cryptology - CRYPTO 1988*, Vol. 403, pp. 27–35 (1990).

[4] Blakley, G. and Meadows, C.: Security of Ramp Schemes, *Advances in Cryptology Lecture Note in Computer Science*, Vol. 196, pp. 242–268 (1985).

[5] Domingo-Ferrer, J. and Mateo-Sanz, J.: Practical Data-oriented Microaggregation for Statistical Disclosure Control, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 14, No. 1, pp. 189–201 (2002).

[6] Drosatos, G. and Efraimidis, P.: Privacy-Preserving Statistical Analysis on Ubiquitous Health Data, *In Proc. of the 8th International Conference TrustBus 2011*, pp. 24–36 (2011).

[7] Fujimori, K.: Practical Use of E-Claim Data for Regional Healthcare Planning (in Japanese), *Japanese Journal of Hygiene*, Vol. 67, pp. 56–61 (2007).

[8] Fushimi, K. and Matsuda, S.: Health Resource Reallocation by Casemix Data in Japan, *BMC Health Services Research*, Vol. 9, No. 1, p. A10 (2009).

[9] Gennaro, R., Rabin, M. and Rabin, T.: Simplified VSS and Fast-track Multiparty Cmputations with Applications to Threshold Cryptography, *In Proc. of the 17th ACM Symposium on Principles of Distributed Computing*, pp. 101–111 (1998).

[10] Goldreich, O., Micali, S. and Widgerson, A.: How to Play Any Mental Game or a Completeness Theorem for Protocols with Honest Majority, *In Proc. of the 19th ACM Symposium on the Theory of Computing*, pp. 218–229 (1987).

[11] Health Level Seven International: HL7 version 3. http://www.hl7.org/.

[12] Jagannathan, G. and Wright, R.: Privacy-preserving Distributed k-means Clustering over Arbitrarily Partitioned Data, *In Proc. of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 593–599 (2005).

[13] Japan Medical Data Center Co., Ltd.: . http://www.jmdc.co.jp/jp/index.html.

[14] Kumakawa, T., Otsubo, K., Hiratsuka, Y. and Okamoto, E.: (Review) Evaluation of Demand for Medical Care Services, Quality of Health Care and Health Policy by Using Electronic Claims Data, *Journal of National Health Institute of Public Health*, Vol. 62, No. 1, pp. 3–12 (2013).

[15] Naser, T., Alhaji, R. and Ridley, M.: Two-Way Mapping between Object-Oriented Databases and XML, *Informatica*, Vol. 33, No. 3, pp. 297–308 (2009).

[16] Sakuma, J. and Kobayashi, S.: Privacy-Preserving Data Mining (in Japanese), *Journal of Japanese Society for Artificial Intelligence*, Vol. 24, No. 2, pp. 283–294 (2009).

[17] Shamir, A.: How to Share a Secret, *Communications of the ACM*, Vol. 22, No. 11, pp. 612–613 (1979).

[18] Sweeney, L.: k-Anonymity: A Model for Protecting Privacy, *International Journal on Uncertainty, Fuzzies and Knowledge-based Systems*, Vol. 10, No. 5, pp. 557–570 (2002).

[19] The MedXML Consortium Non-Profit Organization: Medical Markup Language Specification Version 3.0 (2003). http://www.medxml.net/worldwide/index.htm.

[20] Xin, H. and Weibin, C.: Common Data Mapping System Between XML and Relational Database, *The 2nd International Conference on Computer Application and System Modeling*, pp. 1181–1184 (2012).

[21] Zaidi, S.: Distributed Data Mining from Heterogeneous Healthcare Data Repositories: Towards an Intelligent Agent-based Framework, *In Proc. of the 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002)*, pp. 339–342 (2002).