公募シンポジウム
シンポジウム4
情報の安全な利活用とセキュリティ基盤技術
2021年11月20日(土) 09:10 〜 11:10 C会場 (2号館1階211)

[3-C-1-02] プライバシを保護した分散医療データ統合セキュリティシステム
Privacy-preserving Distributed Medical Data Integration Security System

*宮地 充子[1,2]、高野 祐輝[1]、中正 和久[3]、渡邉 要[4]、成松 宏人[4]（1. 大阪大学大学院工学研究科、2. 北陸先端科学技術大学院大学、3. 山口大学、4. 神奈川県立がんセンター）
*Atsuko Miyaji[1,2], Yuki Takano[1], Kazuhisa Nakasho[3], Kaname Watanabe[4], Hiroto Narimatsu[4]（1. Graduate School of Engineering, Osaka University, 2. JAIST, 3. Yamaguchi University, 4. Kanagawa Cancer Center）
キーワード：big data, PSI, PDDI, privacy-preserving classification protocol

ビッグデータの解析結果は新製品開発など様々な活用が期待され，ビッグデータ解析・利用の促進は重要である．医療分野においては，患者のプライバシを確保しつつ，患者の医療情報を医療分野の発展に利活用できることが望まれる．さらに，データ所有者である患者がデータの利活用に合意できる枠組みの構築が必須である．ビッグデータに関する既存研究では，ビッグデータの効率的な解析手法を改良する研究が多い．一方，本研究課題ではデータ所有者（医療の場合は患者に相当する）に着目し，データのプライバシ保護を実現しつつ，データ解析を実現することを目標とする．[MOSFH16， MNK17, MNK18, MNK19, MTK20]では耐サイバー攻撃の観点からの医療データの安全な管理方法，秘匿分類プロトコル，分散管理されるデータを分散環境のまま，プライバシ情報を削除した医療データのみを統合する方法 PDDI (Privacy-Preserving Distributed Data Integration)について紹介した．
疫学的調査では，異なる機関で管理される同一の患者のデータ突合の重要性は非常に高い．実際，患者は同一の病院に通わずに，複数の病院に通うことが多く，患者の医療データは分散管理される．例えば，癌治癒後，次の病気になった際に異なる病院で治療を受けるケースが多い．また，保健所に管理されるがん検診のデータと癌センターで収集されるがん患者のデータを分散管理したまま，一機関統合することなく，プライバシ情報のない医療データのみが統合できると健診と癌疾患の関係が明らかになる．しかしながら，実際の各データにおいてはデータの表記の揺れなど，データ利活用は容易ではない．．
本稿では， PDDIシステム，保健所の健診データとがん患者データへの統合に向けた PDDI適用， PDDIの並列化実装による高速化について報告する．

プライバシを保護した分散医療データ統合セキュリティシステム

宮地充子*1*2, 高野祐輝*1, 中正和久*3 渡邉 要*4, 成松 宏人*4
*1 大阪大学, *2 北陸先端科学技術大学院大学, *3 山口大学, *4 神奈川県立がんセンター

Privacy-preserving Distributed Medical Data Integration Security System

Atsuko Miyaji*1*2, Yuki Takano*1, Kazuhisa Nakasho*3
*1 Osaka University, *2 Japan Advanced Institute of Science and Technology, *3 Yamaguchi University

It is an important issue to match and analyze medical records managed at different medical institutions independently without leakage of privacy information such as the patient's name, address, etc. In this paper, we introduce basic technologies to realize them. The first technology, Privacy-preserving Distributed Data Integration (PDDI), is a data integration system that enables to match and extract records managed by multiple organizations while preserving privacy.

Keywords: big data, PSI, PDDI, privacy-preserving classification protocol

1 はじめに

ビッグデータの解析結果は新製品開発など様々な活用が期待され, ビッグデータ解析・利用の促進は重要である. 医療分野においては, 患者のプライバシを確保しつつ, カルテ情報を医療分野の発展に利活用できることが望まれる. さらに, データ所有者である患者がデータの利活用に合意できる枠組みの構築が必須である.

ビッグデータに関する既存研究では, ビッグデータの効率的な解析手法を改良する研究が多い. 一方, 本研究課題ではデータ所有者(医療の場合は患者に相当する)に着目し, データのプライバシ保護を実現しつつ, データ所有者, 解析, その利用という 3 つの機能を信頼の環で連結することを目標とする. 参考文献 1)では耐サイバー攻撃の観点から医療データなどの安全な管理方法について提案し, 参考文献 2)では医療データの安全な利活用を促進する 2 技術であるプライバシ保護付きデータ突合手法(PDDI, Privacy Preserving Data Intersection)と秘匿分類プロトコルを報告し, 参考文献 3)では PDDI のソフト評価及び社会実装への展開について方報告した.

PDDI は, プライバシを保護しつつ共通データをキーとしてデータを統合するが, 医療データでは頻繁に必要となる. 実際, 患者は同一の病院に通わずに, 複数の病院に通うことが多く, 例えば, がん治癒後, 次の病気になった際に異なる病院で治療を受けるケースが多い. このため, 疫学的調査では, 異なる機関で管理される同一の患者のデータ突合の重要性は非常に高い [4]. PDDI は異なる機関が保管するデータのうち各機関が保管する共通データをそれ以外の情報は漏らさずに求める手法である.

本稿では, PDDI の方法, 医療機関での利用を想定したシステム設計, ユーザビリティについて述べる. また, がん検診受診の有無とがんの進行度の関係の調査に向けて, 健診・検診機関や自治体のがん検診データとがん登録によるがんり患のデータに PDDI を適用した結果を報告する.

本稿の構成は次の通りである. 第 2 章では共通データをキーとしてデータを統合する方法(PDDI)とプライバシ情報の取り扱いについて述べる. 第 3 章で PDDI のシステム設計, ユーザの利用しやすさについて述べる. 第4章ではがん検診データとがん登録によるがんり患のデータへの適用実験を述べて, 最後に結論をまとめる.

2 PDDI システム

2.1プライバシを保護した集合演算 (PSI)

我々を取り巻く情報社会では, 多種多様なデータが多機関で収集される. 例えば, 小学校で児童がブランコでけがをした事例を考える. このとき, 事故が起こった遊具に関するデータは学校, 病院への救急搬送データは消防署, 傷害・後遺症に関するデータは病院に管理される. つまり, 学校での生徒の事故に関する情報では, 学校, 消防署, 病院がそれぞれ同じ事故で異なるデータを管理する.

学校における事故の予防安全の実現には, 事故の統計的因果モデルの作成が重要である. これにはこのように異なる機関に分散した関連データの統合が必須である. つまり, 異なる機関が独立に収集したデータから生徒の名前などの機微情報は洩れることなく, 同じ生徒の事故の情報を突合(分散多機関データ突合)できると, 事故の詳細なデータの収集が可能になる.

ここで, 異なる機関がもつ医療データの突合方法とプライバシの関係について考える. 単純な方法は, 1つの医療機関が全データを別の医療機関に渡せば, 同じ患者を検索することで, データを突合することができる. しかし, この場合, 本来もつはずでなかった患者の情報である名前, 住所などの機微情報を別の医療機関が入手することになる. 別の方法として, 第 3 の機関(データ預託機関)にそれぞれの病院が医療情報を渡し, その第 3 の機関で突合することもできる. しかしこの場合には, 第 3 の機関に患者の機微情報が移動することになる. つまり, 単純な突合方法は突合に用いる情報が必要となるため, 突合を実施する機関に機微情報が移動し, プライバシ保護を実現することが困難になる.

そこで相互の持つ重要な情報を外部に漏らすことなく, 必要な情報のみのやり取りを行うための有効な技術が必要である. 近年 Private Set Intersection Protocol(PSI) と呼ばれる各機関が持つデータの積集合などの集合演算を, プライバシを保護しつつ実現するプロトコルが注目されている. 典型的な手法である既存のプロトコル [5]では全参加者の入力データ数を一致させなければならないという制限やデータサイズや機関数に依存する処理時間, 通信量が大きな課題となる. 参考文献 6)においてはデータを所有する参加者以外に第三者機関(データ預託機関)を導入する必要があり, その機関もデータを入手する. また, PSI の応用として, 共通データを抽出後, データの加算結果などの処理結果を出力する手法 [7]が提案されている. しかしながら, どの既存方法も, 単一の属性を前

提としており，複数の属性を統合することができない．

本研究室で提案したプライバシを保護した分散データ統合システム Privacy-preserving Distributed Data Integration (PDDI)は突合する属性とそれに付随する解析属性を分けて，突合属性を共有する機関の解析属性のデータ統合を行う方式である．機微情報を他の機関に移動することなくデータ突合を実現する．具体的には，本方式を用いると，データベースから突合する属性と統合する属性を選択し，突合属性をPSI を利用して突合し，突合された属性と付随する必要属性を統合することが可能である．

本方式の持つ特徴について列挙する．

1. どの機関も共通に含まれるユーザ以外の情報について何も入手することはない．(Query ベースとは異なる)

2. 突合に利用する情報は，計算機サーバを含めて，どの機関にも移動しない．

3. 各機関の処理時間は機関数に依存しない．各機関のデータ数に制限や条件はない．

4. 第三者機関によるデータ収集・管理が不要で，容易に導入可能．

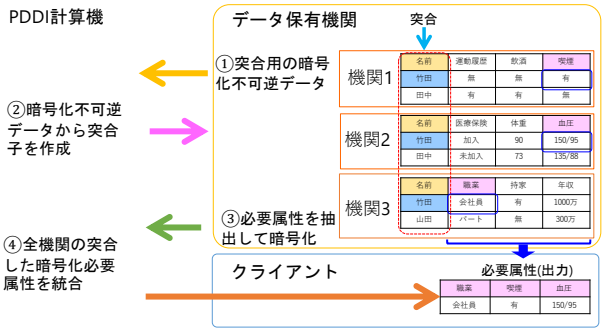

図 1 全体アルゴリズム

2.2 PDDI アルゴリズム

ユーザは情報漏洩を懸念することなく，複数機関が所有するデータの統合を実現する PDDI アルゴリズムについて述べる．図 1 はアルゴリズムの全体像を表す．全機関に存在する同一ユーザを各機関で突合し，そのユーザに関する各機関で保管される情報のみを統合する．

ステップ①

各データ集合の突合したいデータ x（図 2 の場合，名前）をハッシュ関数 H 等で圧縮する．例えば機関 1 のデータの名前 「竹田」は

竹田 → H(竹田)

と一意に不可逆な情報に変換される．その後，さらに準同型暗号 E で暗号化する．

H(竹田) → E(H(竹田))

同様の処理を機関 2，3 で行い，暗号化データを PDDI 計算機サーバに送付する．

ステップ②

準同型暗号は，暗号文の和が元の文の和の暗号文に一致する性質を持つ．この性質を用いて秘匿計算機サーバでは入手した暗号化データをそれぞれ加えることで，各機関のデータの総和の暗号結果を計算できる．図 2 の場合，下記のように，各機関で計算された E (H(竹田))のような名前の暗号文は秘匿計算機サーバで加えられ，名前の和の暗号文となる．

E(H(竹田))+ E(H(田中))+ E(H(竹田))+ E(H(田中))
+ E(H(竹田)) + E(H(山田))
= E(3H(竹田) + 2H(田中) + H(山田))

この暗号化データの総和を各機関に送付する．

ステップ③

受信した各機関のデータの総和の暗号文を復号して，突合したデータ，図 2 の場合，「竹田」が突合属性であることをわかる．次に，各機関に要求される「竹田」の必要属性のみを抽出し，暗号化して PDDI 計算機サーバに送付する．サーバは各機関から送付された暗号化データをシャッフルして連結し，どの機関からのデータであるかを秘匿して，クライアントに送付する．

2. 3 PDDI の実装環境及び環境構築

本節では PDDI ソフトウェアの実装と実験環境について述べる．以下に，PDDI ソフトウェアの設計方針を述べる．

設計方針

1. PDDI ソフトウェアでは核となるビジネスロジックのみ実装すること
2. 設定はすべて設定ファイルで行えること
3. デプロイメントは極力自動で行えるようにすること
4. Web による連携を用意にするために，Restful API を用いて通信を行うこと
5. 通信は極力まとめて行うこと

ビジネスロジックのみを実装することで，PDDI を利用する機関は各々にカスタマイズして利用することが可能となる．しかし，これはつまり，認証などの PDDI 外のメカニズムは独自に実装する必要があるということになるが，一般的な認証用のソフトウェアを使えば良いため実際的な問題にはならない．また，設定ファイルをファイル変更のみで行えるようにすることで，ソフトウェアデプロイメントに対するハードルを緩和させた．従来 PDDI ソフトウェアでは，複数アプリケーション，ライブラリを用いていたため設定方法が煩雑になってしまっていたが，新規ソフトウェアでは容易に設定することができるようになった．さらに，実験環境の構築は Docker を用いて行えるようにしたため，本番環境でも Docker を用いたデプロイが行えるようになる．そのため，デプロイメント時に設定すべき項目が少ない特徴をもつ．

ネットワーク通信のインターフェース部分は Restful API を定義し，本 API 経由で通信を行うようにした．これにより，他のウェブアプリケーションとの連携が容易かつ，ネットワーク利用時のデバッグも容易である．極力データをまとめて送受信することで，ネットワーク利用時のパフォーマンスを改善した．本実装は Python 言語で実現されており，ソースコードレベルでのメンテナンス性もよい．一方，暗号化部分では Python 実装で低速になるので，より詳細なパフォーマンス評価を行う予定である．

以下では実験環境の構築方法について述べる．PDDI ソフトウェアでは Docker と Docker Compose を用いているため，実行するための環境は基本的に以下の 2 回で終了する．

```
$ docker-compose build
$ docker-compose up -d
```

コンテナ起動後に docker コマンドで確認により，以下のように 4 つのコンテナが動作していることがわかる．

```
$ docker ps
CONTAINER ID        IMAGE
a59c5548665f        docker_party2
2f7cb5081da1        docker_dealer
620438d4ecd1        docker_party1
8f067f27bd6f        docker_client
```

ここで，IMAGE が docker_party がデータ提供機関，docker_dealer がデータ統合の補助機関，docker_client が統合データ利用機関である．コンテナ起動後は，コンテナにアタッチして実際に PDDI を動作させることができる．例えば，docker_dealer によりデータ統合用機関を実行するには以下のようにする．

```
$ docker attach pddi-dealer
root@dealer #> python3 /pddi/server.py
 * Running on http://0.0.0.0:5000/
```

従来ソフトウェアではデータ統合用機関用のソフトウェアを実行させることすらハードルが高かったが，今回ソフトウェアではこのように非常に簡単に実行させることが可能となった．

2.3 PDDI ユーザビリティ

本システムでは以下の指針に基づいてユーザーインターフェイスを設計し，ユーザビリティの向上を測った．

直感的な操作：本システムの利用者は必ずしも ICT 技術に精通しているとは限らないため，初期設定を除くほぼ全ての操作がマウスクリックのみで完結するように設計されている．また，利用者がマニュアルに目を通さずとも画面遷移を追うことによって，自然に操作の流れを理解できるようなユーザーインターフェース設計に努めた．

エラーの未然防止：本システムは医療情報など厳重なセキュリティが要求されるデータを利用する．このため，機微情報を他のデータ保有機関へ誤って公開するなどのヒューマンエラーを予防するために，ユーザビリティ上の制約を設けている．まずデータ保有機関のアプリケーションは，他のデータベースに直接接続することができない．データ保有機関の利用者は，アプリケーションが有する独自のデータベースに CSV 形式のデータを登録する．また，照合および統合に利用する属性は，データ保有機関が明示的に許可を与えなければならない．

複雑性の排除：オンライン決済サービス等と同様に，動作原理を理解しなくとも安全性を享受できるように設計されており，利用者がどのような暗号方式によってシステムが実現されているかを意識することはない．このため今後システムの性能改善によって暗号方式が変更されたとしても操作手順には影響しない．暗号化処理やハッシュ化に利用されるパラメータなど専門知識を要する設定はシステム管理者だけがアクセス可能となっている．

導入容易性：PDDI システムの特徴の一つであるスケーラビリティ，すなわち機関数の増加に対して頑強な特性は，処理時間だけではなく，導入容易性においても実現されている．新たにデータ保有機関を追加するときには，秘匿計算機サーバのアドレスと管理者から発行されるパスワード情報のみが必要で，他のデータ保有機関との調整等の複雑な処理は一切不要である．また，本システムで利用されるデータ突合方式はデータ預託機関が不要であり，秘匿計算機サーバをどこに設置しても安全性が保証される．

3　プライバシを保護した疫学調査への応用

PDDI 技術の適用事例として，がん検診データとがん登録データへの適用について検討と効果について検討する．がん検診データは実施医療機関や実施主体（自治体など）で

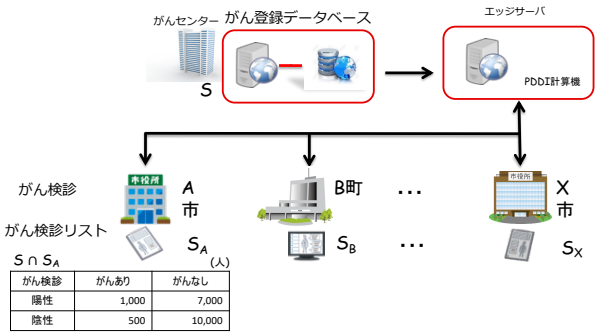

図2　がん検診データとがん登録データへの適用

管理される一方，り患や治療，死亡に関するデータはがん登録の形で各都道府県において管理される．がん検診の精度を評価する場合，検診を受けた個人のり患情報が必要となることから，がん検診データとがん登録のデータの突合による同一人物判定が重要になる．PDDIでは各機関のデータベースは各機関が保管し，同一人物のデータをその個人情報は各機関内から収集することなく，突合する．本事例では，がん検診受診有無とがんの進行度を，それぞれの機関がもつ個人情報はそれぞれの機関内に保持したまま，出力する．図 2 はPDDIの適用事例である．しかしながら，実際の各データにおいてはデータの誤りや表記揺れなどにより，突合による利活用は容易ではない．

本章では，健診・検診データのプライバシを保護した統合を目指して，医療データベースで頻繁に起こる記載の揺れを想定し，統合した実証結果について報告する．

3.1 疫学調査におけるデータ統合の条件

疫学調査に用いられるデータベースにおいて共通の個人識別 ID を有するものはほぼなく，突合に用いるキー情報には氏名，性別，生年月日，住所等の基本情報や，それらをもとに作成された ID を用いることが多い．マイナンバーや医療 ID 等の利活用は現時点では計画段階である [8),9)]．

疫学調査においては，調査対象者のアウトカムを正確に記録する必要があることから同一性の保持が重視される．例えば本章で扱うがん検診精度評価では，検診の判定情報とその後のり患情報が正確に紐付けられずにり患が見逃された場合（偽陰性），感度の過大評価および特異度の過小評価が起こる．また，別人のり患を誤って計上した場合（偽陽性）には逆のバイアスをもたらす．同一性保持のため，ソフトウェアによる突合において，最終的に人力による目視判別を併用することを前提に，偽陽性率が多少高くとも低い偽陰性率を目指す戦略がとられることも多いが，ソフトウェアのみによって突合を行う場合には本技術を活用する疫学研究におけるその手法の種類により戦略は異なる．本章で述べるがん検診精度評価や横断研究の場合には偽陽性，偽陰性いずれにも配慮しバランスの取れた突合方法を選択するのが望ましいが，コホート研究のように継続的な追跡調査を実施する手法の場合には，アウトカムの偽陽性が生じると研究計画によってはその時点で追跡が終了してしまうため，偽陽性率の低減により注意を払う必要があるだろう．

3.2 実験データベースについて

実データで発生する誤りや表記揺れを考慮した実験データベースを構築した．以下に生じうる例を挙げる．まず，データ入力作業において発生するものとして，タイピングや変換のミス，漢字の読み間違い，脱字，記入漏れといった単純な誤りが挙げられる．データそれ自体に起因するものとしては異体字や旧仮名遣いといった文化・制度の変遷に伴う表記揺れがある．データベースや入力者によって記法が異なることによっても表記揺れが発生する．また転居による住所変更，婚姻等に伴う改姓や改名，在日外国人の本名と通称など，突合のキーとなる情報自体が変化する場合がある[9]．

本実験のため，大腸がんと乳がんにおけるがん検診データおよびがん登録データを模して 4 つのデータセットを作成した．突合のキーとなる氏名，性別，生年月日，住所はオンラインテストデータ生成サービスを利用し疑似データを作成した[10-12]．作成した疑似データから大腸がんは 60 件，乳がんは 61 件を抽出して同一人物として扱うこととし，がん検診，がん登録の両方に共通して入力した後，人為的に誤りや表記揺れを追加した．これに重複しないデータを加え，最終的に大腸がん検診 2,000 件，大腸がんり患 17,866 件，乳がん検診 1,048 件，乳がんり患 29,949 件の疑似データセットを作成した．表 1 は追加された誤りと表記揺れの詳細である．

区分	内容	キー	件数	
			大腸がん	乳がん
入力作業に起因するもの	タイピングミス	氏名	3	1
		生年月日	15	0
		住所	6	2
		性別	5	0
	変換ミス	氏名	5	6
		住所	2	0
	読み間違い	氏名	10	8
	脱字	氏名	2	1
	記入漏れ	住所	4	0
	読み仮名のみ	氏名	10	1
表記揺れ	異体字	氏名	7	4
	フォーマット	住所	5	15
データの変化	改姓	氏名	2	1
	本名・通称	氏名	2	0
	転居	住所	2	8
合計（重複あり）			51	36

表 1 データに含まれた誤り・表記揺れ（重複あり）

3.3 PDDI を用いた統合結果

本節では，PDDI を用いて実験データベースの突合を行った結果を示す．実験データベースには個人を特定する情報として以下の 6 つの情報を利用できる．

1. 姓（漢字表記）
2. 名（漢字表記）
3. 姓（カナ表記）
4. 名（カナ表記）
5. 生年月日
6. 性別

本実験ではこれらの 2 個以上の組み合わせについて突合を行った．組み合わせの数は組み合わせの式から計算でき，${}_2C_6+{}_3C_6+{}_4C_6+{}_5C_6+{}_6C_6$ の 57 通りあり，その結果の偽陽性数と偽陰性数について議論する．なお，事前実験より，1 つのみのデータを突合に用いた場合は，著しく結果が悪いことが判明したため本論文では割愛する．

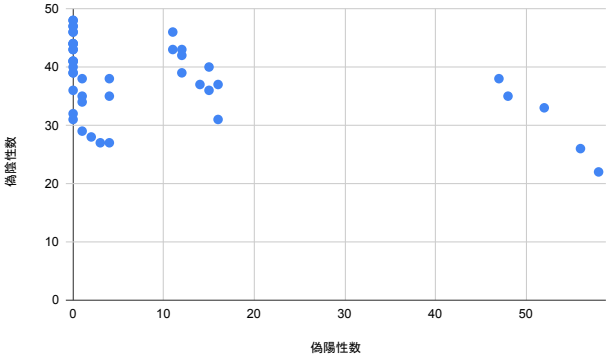

図 3. PDDI を用いた大腸がんデータ突合結果

図 3 は，PDDI を用いて大腸がんの疑似データを，様々な情報の組み合わせを用いて突合を行った結果となる．この図は，縦軸が偽陰性数，横軸が偽陽性数で，それぞれの組み合わせを用いた場合の偽陰性数と偽陽性数を示している．つまり，値が小さい方が精度の良い組み合わせとなる．

突合した結果，偽陰性数の最小値は 27 で，偽陽性数の最小値は 0 となった．これは，3.2 節で示したように，最終的に 60 件のデータが出力されるのが望ましいが，多くても 60-27=33 だけ正しく出力されることを示している．

突合キー	偽陽性数	偽陰性数
生年月日，名（カナ）	3	27
生年月日，姓（カナ），名（カナ）	0	31
生年月日，性別，名（カナ）	2	28
生年月日，性別，姓（カナ）	1	29

表 2．突合結果抜粋

表 2 は，偽陽性数，偽陰性数が特に小さな突合キーを抜粋したものとなる．この表から分かるように，すべてに生年月日が出現しており，この疑似データの場合生年月日が特に有効であることが推測できる．

現在は，大腸がんの疑似データにのみ PDDI を適用したが，今後は乳がんの疑似データや他のデータにも適用して行く予定である．

3.4 実運用に向けて

PDDI システムの特徴は同一人物の異なるデータの統合を分散データのまま実施することにある．今回，がん健診データとがん登録データの突合を実施したが，これは 2.1 章に記載した PDDI のアルゴリズムのステップ①と②を適用した状況である．今後，ステップ③と④の実施に向けて，ん健診データで保管されるがんのステージ情報やがん登録データのがん特定日などの情報を統合することを実現する．これにより，ステップ①と②で存在する偽陽性，偽陰性の取り扱い方法も異なり，より実用に向けてのカスタマイズが可能になると考える．

4 並列化とパフォーマンス評価

本節では PDDI システムの並列化及びパフォーマンス評価を行う．PDDI ではブルームフィルタと呼ばれるデータ

配列の要素ごとに暗号処理が施されるが，この処理に実行時間の 90%以上が費やされる．配列の各要素の暗号処理は互いに独立しており，並列化することは難しくない．今回の並列化実装には Python の multiprocessing モジュールを用いた．

性能評価実験はプロセス数と高速化率の関係を明らかにすることを目的とし，2^{10}, 2^{12}, 2^{14} の各入力データサイズに対して評価する．実験に用いた PC 環境は CPU: Intel(R) Xeon(R) CPU E5-2690 v4@2.60GHz, (28 コア)，Memory 48GB で，一つの PC 上で全機関のプログラムを実行した．図 4 はプロセス数と実行時間の関係を示す．マルチプロセスによる並列化は，プロセスの起動やプロセス間通信にオーバーヘッドが生じるため，高速化の頭打ちとなるプロセス数が存在する．今回は並列化フェーズに関わらず，プロセス数が 16 のときに最も高速化されることが了解される．

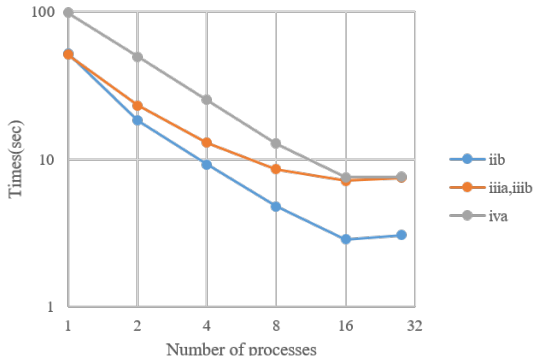

図 4. プロセス数と実行時間の関係

以下では，プロセス数 16 での高速化率について論ずる．図 5 はデータ保有機関で実行されるプログラムの並列化による高速化率を示す．機関数 4，データ数 2^{14} の状況下では，理想値の 81%となる 13.1 倍の高速化率を達成した．機関数やデータ数が少ない状況では高速化率は伸び悩んでいるが，これはマルチプロセス化によるオーバーヘッドのほうが大きくなったためである．図 6 は PDDI 計算機サーバで実行されるプログラムの並列化による高速化率を示す．こちらはデータ保有機関で実行されるプログラムほど効率的な高速化には至らず，理想値の 29%～44%である 4.6 倍～7.0 倍の高速化率となった．これは，データ保有機関プログラムで主となる計算が累乗演算であるのに対し，PDDI 計算機プログラムで主となる計算は入力データ数が多い積演算であることに起因する．PDDI 計算機サーバの存在意義は，時間のかかる処理を外部委託してデータ保有機関の計算負荷を軽減することにあるため，このように並列化の効果が小さい計算処理を PDDI 計算機サーバが請け負うことは理にかなっている．

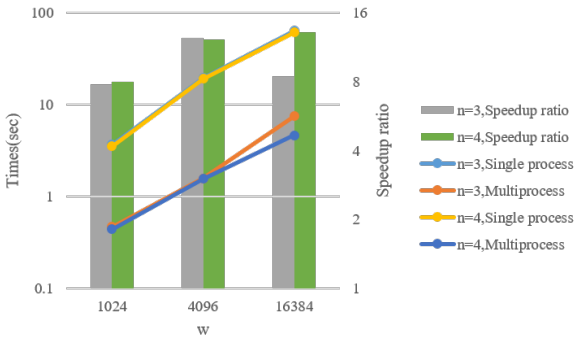

図 5. データ保有機関プログラムの並列化による高速化率

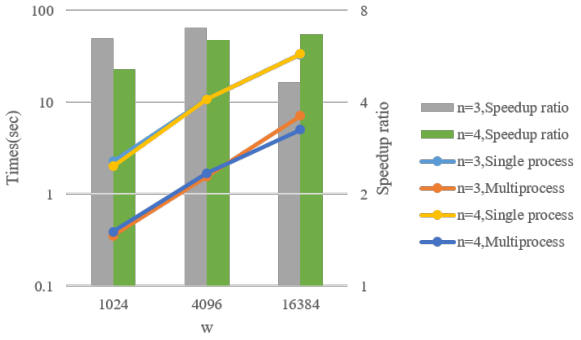

図 6. PDDI 計算機プログラムの並列化による高速化率

5 まとめ

本稿ではプライバシを保護しつつ医療データの利活用を可能にするプライバシを保護した分散データ統合システム(PDDI)について紹介し，疫学調査への応用，さらには並列化によるパフォーマンスについて紹介した．本技術を用いると，診療所や病院などの複数機関で分散管理されるデータを，第三機関を用いずに，互いに非開示のまま共有データに関連する情報のみを突合することが可能である．医療現場においては，関連するデータが複数の異なる機関で管理されるケースが頻繁に起こる．例えば，同じ患者が異なる疾患にかかった場合，各疾患を専門とする複数の病院に通うことが考えられる．このように独立した 2 つの医療機関で管理された異なる疾患同士には因果関係がある可能性がある．この時，同一の患者のデータを，それぞれの機関がもつ個人情報はそれぞれの機関内に保持したまま，必要な医療データのみ突合できると，異なる病気の因果関係に関する詳細なデータの収集が可能になる．PDDI では，個人を特定できる情報はどの機関にも移動せずに突合とデータ抽出を可能にし，突合を行う医療機関数が増えても，高速に処理できる．今後，各種医療データへの利活用により，異なる病気の因果関係の究明や，稀少疾患の治療方法の発展への貢献が望まれる．

謝辞

本研究の一部は文部科学省の平成 30 年度「Society 5.0 実現化研究拠点支援事業」さらに JSPS 科研費 JP21H034438 の助成を受けています．

参考文献

1) 宮地充子，面和成，蘇春華，布田裕一，波多野哲也，西田昌平.ビックデータ統合利活用促進のためのセキュリティ基盤技術. 医療情報学会 2016
2) 宮地充子，中正和久，河内亮周. プライバシを保護した多機関データ突合システムについて. 医療情報学会 2017
3) 宮地充子,高野祐輝,中正和久.プライバシを保護した分散医療データ統合セキュリティシステム. 医療情報学会 2020
4) Gon, etc. Validation of an algorithm that determines stroke diagnostic code accuracy in a Japanese hospital-based cancer registry using electronic medical records. BMC, Dec., 2017.
5) L. Kissner and D. Song. Privacy-preserving set operations.CRYPTO 2005, LNCS 3621, 241-257, Springer, 2005.
6) Many, Burkhart, and Dimitropoulos. Fast private set operations with sepia. Technical Report, 345, 2012.
7) Ion, Kreuter, et al. On Deploying Secure Computing Commercially: Private Intersection-Sum Protocols and their Business Applications. IACR Cryptology ePrint Archive: 723, 2019.
8) 医療等情報の連結推進に向けた被保険者番号活用の仕組みについて.医療等情報の連結推進に向けた被保険者番号活用の仕組みに関する検討会報告書. 厚生労働省.
9) 久保慎一郎，野田龍也，明神大也，東野恒之，松居宏樹，加藤源太，今村知明. レセプト情報・特定健診等情報データベース(NDB)の臨床研究における 名寄せの必要性と留意点. 日本健康開発雑誌 p11-10, 38 号, 2017
10) 疑似個人情報データ作成サービス http://hogehoge.tk/personal/
11) なんちゃって個人情報 http://kazina.com/dummy/
12) テストデータジェネレータ http://yamagata.int21h.jp/tool/testdata/