個人ゼロ識別秘匿データ共有分析

菊池 亮[1,a)]　高橋 元[1]　高橋 克巳[1]　五十嵐 大[1]

概要：複数の個人データを組み合わせて分析するデータ共有分析は，データの数や種類を増やすことができるので，より信頼性の高い結果や未知の知見を得ることが期待できる．例えば，医療データと運動データを組み合わせることによる健康寿命の延伸プランの策定などがあり，さらにはパンデミック対応で感染データと位置データの共有の是非が議論されているところである．しかし，個人データの共有には個人の不安につながるリスクがあり，データ共有分析は活発に行われているわけではない．個人データの共有を行うと「大きな」データ，すなわち一人の個人に対して多数の属性情報が結合されたデータが作られるため，個人データの閲覧や，目的外の利用が過度なレベルで起きる可能性がある．もし，分析の開始から終了まで個人がわからないデータ共有分析ができれば，前記課題を解決し，データを保護した活用が可能となり，データ活用が促進すると考えられる．本論文では，個別の技術において議論されてきた事柄を整理・再設計し，個人がわからないデータ共有分析が，秘密計算を用いて実現できることを示し，その要件を整理する．この個人がわからないデータ共有分析を**個人ゼロ識別秘匿データ共有分析**と呼ぶこととする．個人ゼロ識別秘匿データ共有分析はデータ共有分析を保護と活用の両面から支えるという意味において個人情報保護に貢献する．

キーワード：統合分析，プライバシー保護，データ共有，秘密計算，秘密分散

Zero-identification secure data-sharing analysis

Ryo Kikuchi[1,a)]　Gen Takahashi[1]　Katsumi Takahashi[1]　Dai Ikarashi[1]

Abstract: Data sharing analysis, which combines and analyzes multiple personal data, can increase the number and variety of data, and therefore, more reliable results and unknown findings can be obtained. For example, a plan to extend healthy life expectancy can be developed by combining medical and exercise data. The advantages and disadvantages of sharing infection data and location data are still being discussed in the pandemic response. However, data sharing analysis is not actively conducted due to the risk of personal data sharing leading to providers' anxiety. The sharing of personal data creates "big" data, i.e., a large number of attribute information combined for a single person, which can lead to excessive levels of personal data exposure and unintended use. If a data-sharing analysis that prevents individuals from being identified from the beginning to the end of the analysis is possible, it would solve the above problem, enable protected use of data, and promote data utilization. In this paper, we organize and redesign discussion on specific technologies, show that the data sharing analysis, where persons are not identified, can be realized using secure computation, and summarize its requirements as well. We call this data-sharing analysis as *zero-identification secure data-sharing analysis*. This technology contributes to the preservation of personal information in the sense that it supports both privacy protection and data utilization of data-sharing analysis.

Keywords: collaborative analysis, privacy-preserving data mining, secure computation, secret sharing, horizontal federated learning

1. はじめに

複数の個人データを組み合わせて分析するデータ共有分

[1] NTT セキュアプラットフォーム研究所
NTT secure platform laboratories
[a)] kikuchi_ryo@fw.ipsj.or.jp

析は，データの数や種類を増やすことができるので，より信頼性の高い結果や未知の知見を得ることが期待できる．例えば，医療データと運動データを組み合わせることによる健康寿命の延伸プランの策定，小売りデータと移動データを組み合わせることによる街づくりなどでデータ共有分析がより一層重要になり，さらにはパンデミック対応で感染データと位置データの共有の是非が議論されているところである．

一方，データ共有分析は活発に行われているわけではない．個人データの共有には個人情報保護法上の制約（利用目的による制限，第三者提供の制限等）があるが，この制約が意味することも含め，個人の不安につながるリスクがあるためである．リスクは技術的に次のように考えることができる．データ共有を行うと「大きな」データ，すなわち一人の個人に対して多数の属性情報が結合されたデータが作られるため，個人データの閲覧（個人の属性情報が一度に知られてしまう．漏洩はその最悪のケース）や，目的外の利用（思いもしない分析が個人に対してなされる）が過度なレベルで起きる可能性がある．

個人データの共有によって特定の個人の分析をするのであれば，従来の個人情報保護の考え方に従うべきである．一方，個人データの活用にいわゆる統計的な利用（特定の個人との関係が排斥され集団としての傾向や性質を把握すること）が認められていることが知られている．もし，分析の始め（共有）から終わり（統計）まで “個人がわからない” データ共有分析ができるのであれば，データ共有に係る個人の不安を払しょくすることができると考えられる．

そのような個人がわからないデータ共有分析に有用と思われる技術要素として，データの中身がわからないまま分析ができる秘密計算技術や，分析結果からプライバシー情報を保護するアウトプットプライバシーとよばれるものが研究されている．しかし，データ共有分析を行う目的においては，これらの各技術を組み合わせた場合の効果や安全性要件等の総合的な整理がなされていないという課題がある．

1.1 本論文の貢献

本論文では，個人がわからないデータ共有分析に焦点を当て，秘密計算やアウトプットプライバシーの文脈において議論されてきた事柄を整理・再設計し，個人がわからないデータ共有分析が，秘密計算を用いて実現できることを示す．この個人がわからないデータ共有分析を**個人ゼロ識別秘匿データ共有分析**と呼ぶこととし，個人ゼロ識別秘匿データ共有分析を実装するための要件を整理する．

2. 秘密計算による新しいデータ活用方法の提案 – 個人ゼロ識別秘匿データ共有分析

秘密計算では，データを秘密分散データの状態を保って全てのデータ処理を行うため，データの中身がわからないまま分析を行うことができる（秘匿データ共有分析）．データが秘匿されているので，データの登録から分析の過程において個人がわからないまま分析ができるといえる．

さらに，秘密計算は，計算結果を統計情報に限定することができる．なお本論文において統計情報とは，特定の個人との対応関係を排斥した統計を意味するものとする．計算結果が統計情報ならば，誰も計算結果から個人を識別できない．これらのことから，データは登録時から分析終了まで個人がわからない（個人ゼロ識別分析）．以上から個人ゼロ識別秘匿データ共有分析が実現できる．

個人ゼロ識別秘匿データ共有分析を改めて定義すると，複数の個人データを組み合わせて行う個人データ共有分析で，分析の開始から終了までデータが秘匿されていて，分析結果を含む分析の全過程で特定の個人の識別が起きないものである．

以降，3 章にて，「秘密計算の仕組みと秘匿データ共有分析を行える要件」について説明する．4 章にて，「秘密計算の分析結果から個人がわからない仕組みと個人ゼロ識別秘匿データ分析の成立要件」について説明する．

3. 秘密計算の仕組みと秘匿データ共有分析の成立要件

秘密計算では，個人データを秘密分散データとして扱い，処理に秘密計算プロトコルを用いる．秘密分散データがデータの中身がわからない状態を担保し，秘密計算プロトコルがデータの中身がわからないデータ処理を実現する．以下に，秘密計算がどのようにデータの中身がわからない状態でデータ処理を行うかを説明する．なお，秘密計算にはいくつかの方式がある．本稿では，現在総合的な実用性が高いとされる秘密分散を用いた秘密計算に関する説明を行う．

3.1 秘密分散技術

秘密分散技術とはデータを安全に保管するための技術として知られてきたものである．保管するときはデータを秘密分散データに分割し，使うときは必要な数の秘密分散データを集めて復元することで元のデータに戻すことができる．秘密分散データとはそれ単体では何も中身がわからないデータである．従って，個人データを複数の秘密分散データに分割しそれぞれを分離して管理する場合，それぞれの秘密分散データからは特定の個人が識別できない．

次に，秘密分散アルゴリズム，秘密分散データの性質や秘密分散データの安全性要件について説明する．

3.1.1 秘密分散技術に用いるアルゴリズム

秘密分散技術は対応する秘密分散アルゴリズムと復元アルゴリズムの組からなる．秘密分散アルゴリズムは元データを 3 つの秘密分散データに分割して，分離管理する．復

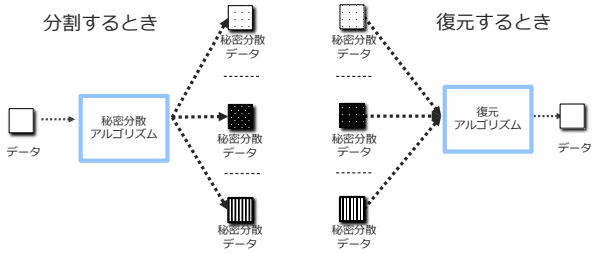

図 1　秘密分散技術の一連の流れ

元アルゴリズムは 3 つの秘密分散データのうちいずれか 2 つの秘密分散データを使い元データを復元する．一連の流れを図 1 に示す．なお，ここでは 3 分散・2 復元の秘密分散で説明するが，分散数・復元数の選び方は 3・2 に限定されない．

この秘密分散技術に用いるアルゴリズムは，ISO/IEC で国際標準化されている [1], [2].

3.1.2　秘密分散データの性質

秘密分散データには以下の性質がある．なお，秘密分散データはシェアと呼ばれることがある．

- 秘密分散データ単体は元データが何なのか全くわからない無意味なデータである．
- 同じデータを再度秘密分散しても，異なる秘密分散データが作られる．
- 秘密分散データの安全性はアルゴリズムが知られていても損なわれることはない（アルゴリズムを隠す必要がない）．

3.1.3　秘密分散データの安全性要件

秘密分散データは以下の要件を満たすとき安全（データの中身がわからない）である．この要件を「秘密分散データの安全性要件」という．

【秘密分散データの安全性要件】
- 秘密分散データを適切なアルゴリズムで作成する．
- それぞれの秘密分散データを分離管理する（同時に作成したそれぞれの秘密分散データを一定数以上まとめずに分離して管理する）．

3.2　秘密計算プロトコル

秘密分散技術を用いた秘密計算とは，データを秘密分散した秘密分散データとして扱い，データの計算を秘密計算プロトコルと呼ばれる手続きに従って，全て一貫して秘密分散データの形式で行うことである．大まかな流れを図 2 に示す．

この際，秘密分散データの安全性要件を満たすことで，秘密計算では，一切データの中身がわからないままデータ処理することができる．計算結果は秘密分散データの形式で出力され，その秘密分散データを復元することで最終的な計算結果が得られる．

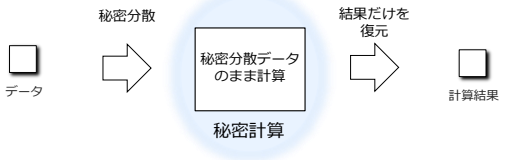

図 2　秘密分散技術を用いた秘密計算

3.2.1　秘密計算プロトコルのデータの流れ

以下に，秘密計算プロトコルのデータの流れを説明する．クライアントは，個人データを保有し，クライアントの数に限りはない．サーバは秘密計算を運用し，サーバ数は 3 が典型的である．図 3 も参照されたい．秘密計算プロトコルのデータの流れは，登録時は (1)，計算時は (2) から (4) となる．

(1) クライアントは個人データを複数の秘密分散データに分割して各サーバに分離して登録する
- 各サーバは登録されたそれぞれの秘密分散データから元データの内容がわからない

(2) クライアントはサーバに計算要求を出す

(3) 各サーバは計算を行いそれぞれの計算結果をクライアントに返す
- 計算過程（計算結果の出力も含む）において各サーバは元データの内容がわからない
- 計算結果も秘密分散データなので各サーバは計算結果の中身がわからない

(4) クライアントは，各サーバから得た秘密分散データである計算結果から，最終的な計算結果を復元して得る

上記 (3) においてサーバが行う計算は下記である．

(i) 各サーバは登録された秘密分散データを用いて所定の個別計算[*1][*2]を行い，その計算結果の値[*3]を（出力せずに）秘密分散して各サーバにそれぞれ分離して登録する（この一連の手続きをラウンドと呼ぶ）

(ii) ラウンドを所定回数繰り返す[*4]

(iii) 所定回数終了後，最終的な計算結果を秘密分散データとしてクライアントに出力する

3.2.2　秘密計算プロトコルの安全性要件

以下の 3 点から秘密計算の各サーバは，元データの中身がわからず，元データの内容を全く知ることができない．

- 秘密計算においてサーバが行う処理は秘密分散データ同士の計算と，その結果を秘密分散することだけである
- 秘密分散データ同士の計算結果は，秘密分散データが秘密分散データの安全性要件を満たしている限り，そ

[*1] 個別計算の内容はラウンドごとに異なる．
[*2] 個別計算の具体的内容は，要求された計算（例えば「平均値を求める」）に対して秘密計算内で割り当てられる．
[*3] 個別計算結果も秘密分散データの性質を持つもので各サーバはその中身がわからない．
[*4] ラウンド回数は，秘密計算内で割り当てられる．

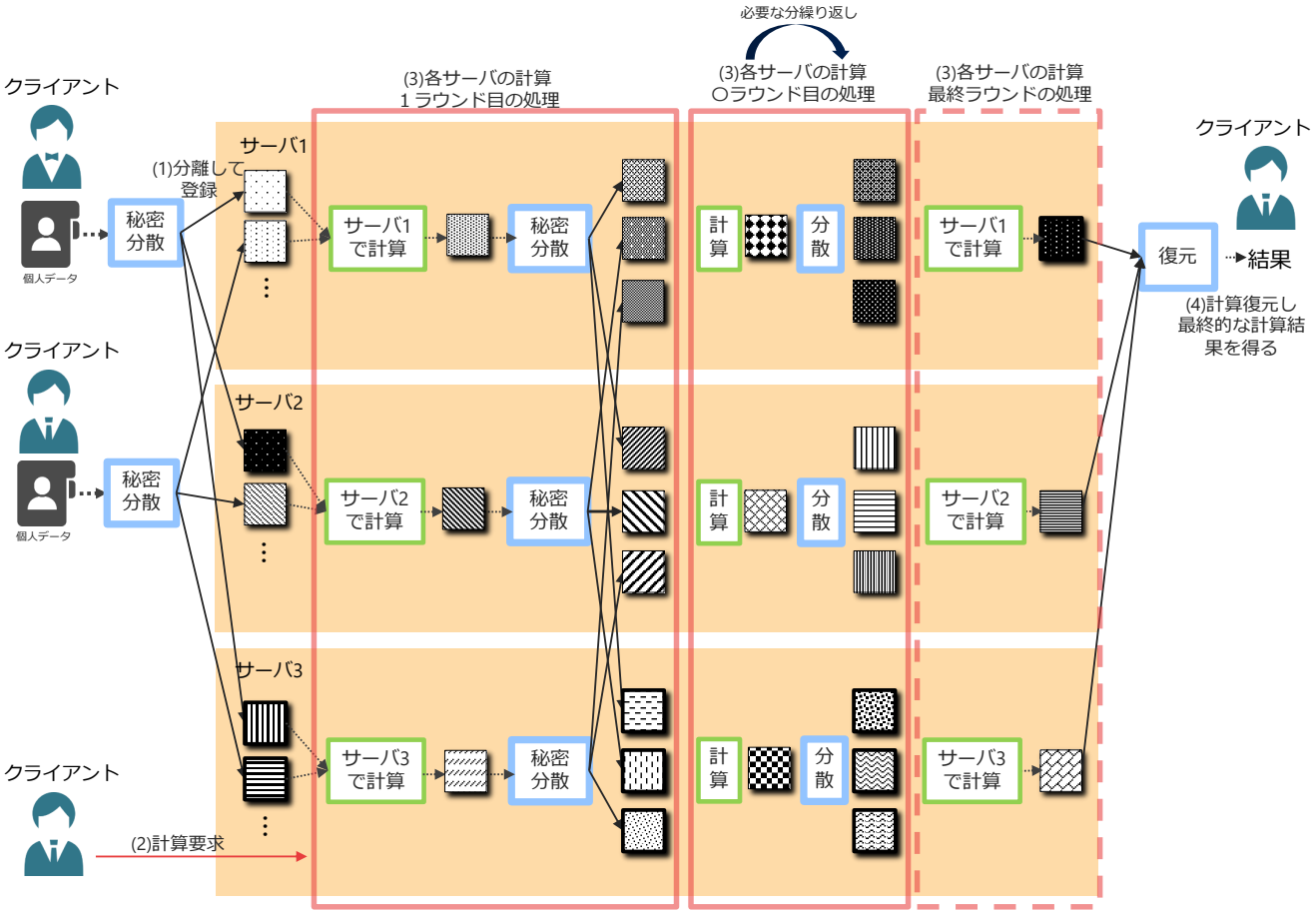

図 3　秘密計算プロトコル

の中身がわからない（秘密分散データの性質を持つ）

- 秘密分散データを秘密分散しても秘密分散の性質を持ち続ける

上記から，秘密計算プロトコルは以下の要件を満たすとき安全（データの中身がわからない）である．この要件を秘密分散プロトコルの安全性要件という．

【秘密計算プロトコルの安全性要件】
各サーバへ入力された秘密分散データおよび計算中の全ての秘密分散データ（秘密分散データの性質を持つ個別計算結果）が，秘密分散データの安全性要件を満たすこと．

3.3　秘匿データ共有分析の成立要件

秘密計算においてデータの中身がわからない状態で分析を行える要件は，全てのデータを秘密分散データとして扱い，全ての計算を秘密計算プロトコルで行い，各々の安全性要件が満たされていることである．

【秘匿データ共有分析の成立要件】

- 全てのデータを安全性要件が満たされた秘密分散データで行う．
- 全ての計算を安全性要件が満たされた秘密計算プロトコルで行う．

【秘密分散データの安全性要件】

- 秘密分散データを適切なアルゴリズムで作成する．
- それぞれの秘密分散データを分離管理する（同時に作成したそれぞれの秘密分散データを一定数以上まとめずに分離して管理する）．

【秘密計算プロトコルの安全性要件】
各サーバへ入力された秘密分散データおよび計算中の全ての秘密分散データ（秘密分散データの性質を持つ個別計算結果）が，秘密分散データの安全性要件を満たすこと．

3.4　（参考）秘密分散データの分離管理について

秘密分散データの分離管理を実現する運用方法に関しては，いくつかの実現方法があると考えられる．データ管理者（Controller）は，共有する個人データの保有者（群）（図 3 ではクライアント）が個人データの管理者である．データ処理者（Processor）を管理者と独立の存在とすることも可能で，本稿および図 3 では管理者と処理者が独立の前提で説明をしている．この場合，秘密分散データの分離管理は，管理者の責任のもと，処理者が行う．

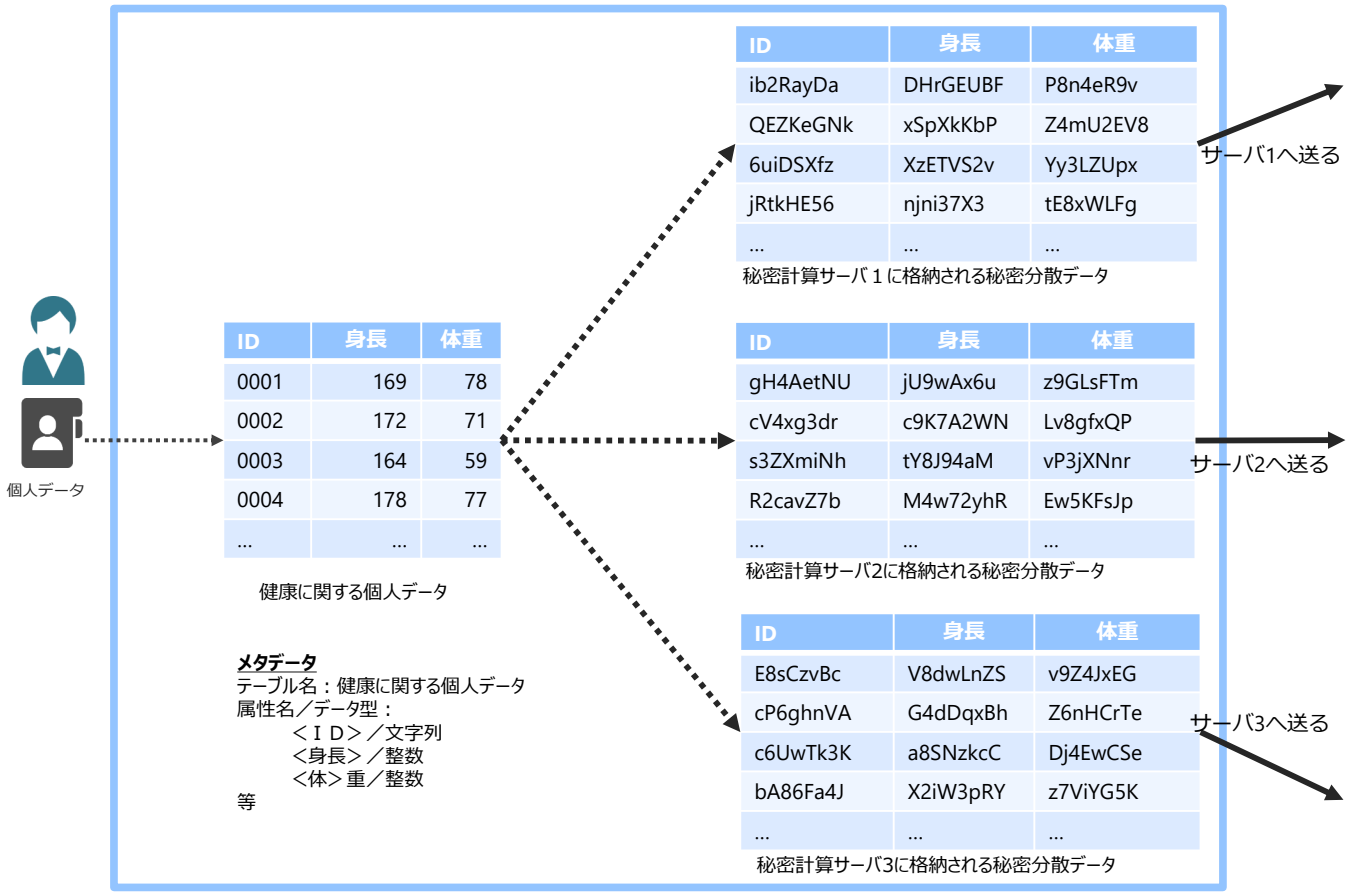

図 4　個人ゼロ識別秘匿データ共有分析のデータ例

4.　秘密計算の分析結果から個人がわからない仕組みと個人ゼロ識別秘匿データ共有分析の成立要件

ここまでで，データの中身がわからないままデータ処理を実現する秘匿データ共有分析について説明した．さらに，秘密計算のデータ分析の結果からも個人がわからないことを説明する．基本的な考え方は次のとおりである．秘密計算は，分析結果を限定することができる．秘密計算の分析結果を特定の個人との対応関係が排斥されている統計情報に限定することで，秘密計算のデータ分析の結果から個人をわからないようにできる．以下に，秘密計算の分析結果を適切な統計情報，すなわち特定の個人との対応関係が排斥されている統計，に限定する方法について説明する．

4.1　秘密計算の分析結果を統計情報に限定する方法

クライアントが統計の作成を依頼（例，平均値，クロス集計値等を想定）した際に，サーバは分析結果が適切な統計情報かどうかを秘密計算の仕組みを用いて判定して，適切な場合のみ作成した統計情報を分析結果として出力する．これにより，秘密計算の分析結果を統計情報に限定できる．適切な基準は，データ管理者（図 3 の場合クライアント）が決定し，その基準があらかじめ秘密計算に正しく実装されておりそれ以外の場合は計算結果の出力をしないものとする．また，その基準の考え方は，統計情報の量的問題（その統計が対応する個人の数）や質的問題（その統計を作成することが個人や公共の利益を損することの有無）を考慮するとし，PIA（プライバシー影響度評価）等の考え方を取り入れてデータ管理者が事前に実施する．

4.2　個人ゼロ識別秘匿データ共有分析で個人がわからない分析を行える要件

秘密計算を用いた個人ゼロ識別秘匿データ共有分析において，個人がわからない状態で分析を行える要件は，以下を満たすことである．

【個人ゼロ識別秘匿データ共有分析の成立要件】
- 秘密計算が【秘匿データ分析の成立要件】を満たした分析を行う．
- 秘密計算が分析結果を判定して，適切な統計情報の場合のみ結果の出力を行う．

4.3　（参考）複数回の分析について

一般に統計情報を計算する際に，その分析が複雑な場合は，データベースに対して何回かの分析を行うことがある．個人ゼロ識別秘匿データ共有分析においても同様のことが必要な場合は，何回かの分析におけるデータ分析の結果も統計に限定する．また，一般に統計計算を統計の要件を変えて複数回数行うと，個人への関係が推定できることが指摘されている．個人ゼロ識別秘匿データ共有分析において

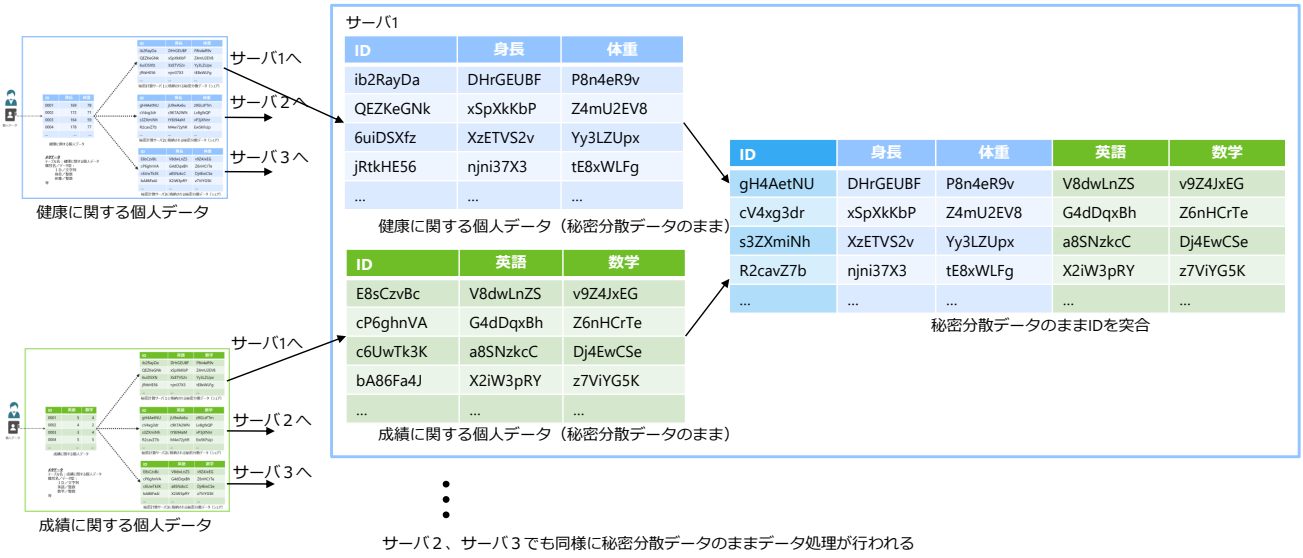

図 5　秘密計算により異なる属性項目を持つデータを突合する例

も同様のリスクは存在するので，統計リクエストの回数を制限するなどの措置をとる．

5. 秘密計算を用いた個人ゼロ識別秘匿データ共有分析の例について

本節では，秘密計算を用いた個人ゼロ識別秘匿データ共有分析の例について説明する．秘密計算が利用するデータの例を図 4 に示す．

この例のデータはテーブル形式のデータベースであり，秘密分散データのままデータ処理される．

このような個人データを秘密分散データとして扱う場合，下記の特徴をもつテーブルとして表現される．

- 個人データは秘密分散データにおいても個人に対応した単位（行単位）で管理される
- 秘密分散データが作成される単位は，個人の属性項目ごとである（レコード数 n，属性項目数 m のデータの場合は $n \times m$）
- 以下のようなメタデータはクライアントからサーバに共有される（秘密としては扱われない）
  - テーブル名（例：健康に関する個人データ）
  - 属性名（例：＜ ID ＞，＜身長＞，＜体重＞，...）
  - データ型（例：文字列，整数，...）
  - レコード数，等

個人ゼロ識別秘匿データ共有分析が想定する利用例は，複数のデータ保有者が持つ個人データを持ち寄り組合せて分析することである．特に，異なる属性項目を持つ同一の個人のデータを共有して，より多項目のデータを分析するケースは，ビッグデータの活用として期待のあるものである．この分析を個人ゼロ識別秘匿データ共有分析ではデータがわからない状態で分析を行い，分析結果から個人がわからないことを満たして行うことができる．一般のデータベースと同様に，メタデータ（テーブル名や属性名，データ型など）が共有されていればデータの共有が可能で，異なる属性項目の同士のデータであっても，なんらかの共通の ID 等があれば，その ID がなんであるか分からず図 5 のように突合して横断的な分析を行うことができる．

なお，秘密計算で突合に用いる ID は，ハッシュ化した ID を突合する場合とは異なる．秘密計算では，平文の時点で同じ ID であっても秘密分散データは秘密分散するごとに違う値になるため，同じ ID が共通にあるのかどうかすらわかることがない．

6. おわりに

本稿では，複数の個人データを分析の開始から終了まで個人がわからないまま共有し分析する，個人ゼロ識別秘匿データ共有分析について提案した．2 章で個人ゼロ識別秘匿データ共有分析の概念について説明した．秘密計算を用いてデータの中身をわからないまま分析できる，秘匿データ共有分析ができる要件，さらに，分析結果を統計情報に限定することで分析結果からも個人がわからない個人ゼロ識別秘匿データ共有分析ができる要件を 3 章および 4 章で説明した．5 節では，個人ゼロ識別秘匿データ共有分析において特に価値が高いと考えられる異なる属性項目を持つ同一の個人のデータを共有して，より多項目のデータを分析する例を説明した．個人情報の活用において，多様なデータを組み合わせて分析することにより新たな価値を発見することに期待が寄せられている．秘密計算による個人ゼロ識別秘匿データ共有分析は，分析の開始から終了まで

個人がわからないデータ共有分析であり，新しい価値で個人情報保護に貢献できると考えられる．それゆえに，この秘密計算を用いた個人ゼロ識別秘匿データ共有分析のあり方に関する議論には十分な意義があると考える．

参考文献

[1] ISO/IEC 19592-1:2016. Information technology – security techniques – secret sharing – part 1: General, 2016.

[2] ISO/IEC 19592-2:2017. Information technology – security techniques – secret sharing – part 2: Fundamental mechanisms, 2017.