様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ８ 日現在

機関番号：１２１０２
研究種目：研究活動スタート支援
研究期間：2019～2022
課題番号：１９Ｋ２４２２２

研究課題名（和文）医療・介護等データベース利用のボトルネック解消を目的とした人工データベースの開発

研究課題名（英文）Developing a method for resolving bottleneck in analyses of national databases

研究代表者
久米　慶太郎（Kume, Keitaro）

筑波大学・医学医療系・助教

研究者番号：７０８５３１９１

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本研究では、多大な価値を有するレセプト情報・特定健診等情報データベース（ＮＤＢ）について、データベースに含まれるデータの有用性を残しつつ、個人のプライバシーを保護するような手法の検証・評価研究を行った。その結果、属性の一般化を行ったりデータの偏りをなくしたりするような手法では、十分なプライバシー保護を確保するためにはデータの有用性が十分に維持されないこと、乱数ノイズを加える手法が有望である可能性が示された。また、あわせて検証用ダミー・データセットの開発を行った。

研究成果の学術的意義や社会的意義
レセプト情報・特定健診等情報データベース（ＮＤＢ）は、多大な価値を有する世界的にも類を見ないデータベースであるが、個人の医療情報という性質上データへのアクセス手段が限定されているというボトルネックが存在するため、その価値に対して利活用が十分には進んでいない。本研究は、プライバシーを保護しつつデータベースの有用性を残すための手法について何が適切なのかという情報を提供することでこれを解消し、社会へＮＤＢの成果を還元するための一端を担うことができると考えている。

研究成果の概要（英文）：National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB) is a very valuable database, but its full potential has not been utilized because of barrier of restricted and limited access to the data due to its importance of health or medical information privacy. To address this problem, we performed experiments for validation and evaluation of methods to protect sensitive data while preserving the utility of the data and tried to determine the suitable method. As a result, it was shown that methods that substituting the attribute value with more general values or eliminate data bias do not sufficiently maintain the usefulness of the data to ensure sufficient privacy protection, and that methods that add random noise may be promising. In addition, we developed a dummy NDB dataset for these experiments.

研究分野：医療情報学

キーワード：医療データベース　プライバシー保護　ビッグデータ　匿名性　医療情報学

１．研究開始当初の背景

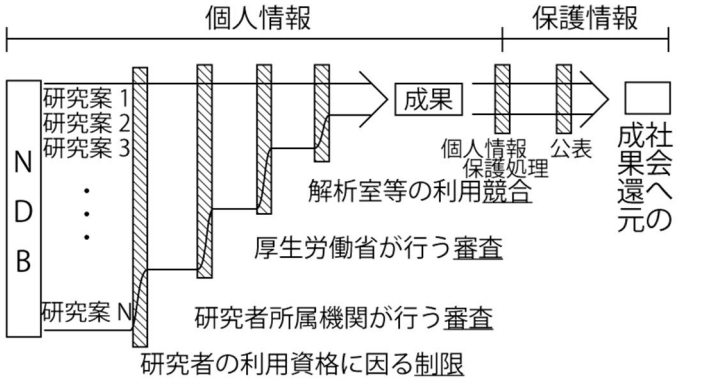

図１．研究開始当初の背景

レセプト情報・特定健診等情報データベース（ＮＤＢ）は、レセプト情報を収集・管理している悉皆性の高い世界的にも類を見ない有用な膨大量のデータ（ビッグデータ）である。近年、ビッグデータからの知識発掘（データマイニング）や、人工知能（ＡＩ）技術の分野が飛躍的な発展をとげており、他の様々な分野において、これらビッグデータ及びＡＩの活用が期待されている。高齢化社会が急速に進展するなか、我が国の医療財政は現在厳しい状況に直面しており、効率のよい医療システムを構築することは喫緊の課題である。そこで、ＮＤＢに格納されているデータをＡＩ等によって分析し、保健医療の質の向上と医療費の適正化を図ることが求められている。

しかし、その目的を達成するためのデータ利用環境は未だ整っていないと考えられる。ＮＤＢに格納されている治療歴等の医療情報は極めてセンシティブな個人情報であり、そのためにＮＤＢの利用には厳格な審査を経る必要がある。このように、極めて有用なデータが存在するにもかかわらず、ＮＤＢへのアクセス手段が限定されているというボトルネックが存在し、活用が進んでいない状況であった。

２．研究の目的

そこで本研究では、「十分なプライバシー保護を提供しつつ情報の損失を抑え、アクセス手段の制約なしに事実上ＮＤＢ情報を利用可能にするような、人工・疑似ＮＤＢ構築やプライバシー保護の方法論の開発と検証」を目的とし、これによってＮＤＢ利活用のボトルネックを解消することを試みた。

３．研究の方法

（１）検証用ダミー・データセットの構築

本来、ＮＤＢのデータセットは厚生労働省有識者会議の厳格な審査による承認等を経た上で利用可能なデータである。しかし、どのような形式でデータが利用可能であるかは、電子レセプトの仕様書の公開情報や厚生労働省が設置する匿名レセプト情報・匿名特定健診等情報の提供に関するページで提供されている。そこで、本研究ではプライバシー保護手法の検討と評価には実際のデータを用いることは必ずしも必要とされないことから、これらの情報をもとにダミー・データセットを構築しプライバシー保護手法の検証を行うこととした。ＮＤＢは巨大なデータセットであるため、効率的な検証を行うために、特に治療のエピソードに注目し、関連のあるレセプトの一部レコードに絞り込んでダミー・データセットを構築することとした。ＮＤＢには、医科レセプト、調剤レセプト、ＤＰＣレセプト、歯科レセプトの情報の他、健診情報が存在するが、本研究では医科レセプト、調剤レセプト、ＤＰＣレセプトの各レコードを対象とした。点数の項目など、他の関連レコードを参照したり外部のマスタを参照したりすることによって復元可能な項目については、本検証においては効率性のために除外した。

（２）プライバシー保護手法の検証

このダミー・データセットを用いて、プライバシー保護手法の検証を行った。プライバシー保護の度合いの評価基準としてｋ匿名性、ｔ近似性を用いた。ただし、これらの基準値には明確なものは定められていないため、順次加工後の結果を分析しながら決定することとした。それを満たすような加工を各レセプトの各レコードの項目(例:医科レセプト RE レコードなら年齢区分、男女区分、都道府県などの項目がある）に対して適用し、それぞれについて元のデータセットの特性や有用性（ユーティリティ）をどれほど維持しているかを検証した。

４．研究成果

（１）成果：検証用ダミー・データセットの構築

医科レセプト、調剤レセプト、ＤＰＣレセプトのレセプト情報について、公的に提供されてい

る公開情報・利用可能データをもとに、200 ID 分のダミー・レコードを生成した。さらにこれらを実際のＮＤＢ研究の際の規模を想定し、機械学習の手法を用いて 10 万 ID 分までデータ拡張を行うことによって、検証用ダミー・データセットの構築に成功した。

（２）成果：プライバシー保護手法の検証

当初はｋ匿名性、ｔ近似性をプライバシー保護の程度の評価指標としたが、大元の（ダミー）データセットの個人の特定可能性が非常に高く、評価基準値を大幅に緩めたとしても、その基準を満たすためにはデータセットの大部分の情報を削除または一般化する必要があることがわかった。そのため、このようなプライバシー保護手法では元データの有用性を十分に維持できないことが分かった。そこで新たなプライバシー保護手法として差分プライバシー手法を検討した。その結果差分プライバシー保護処理によってノイズを追加された後のダミー・データセットの基礎情報の集計・分析値は、保護前のダミー・データセットの集計・分析値に収束することが分かり、元のダミー・データセットの有用性が維持されていることが分かった。これによってＮＤＢにおいても差分プライバシーがプライバシー保護に有用である可能性が示された。本研究の成果は、論文として投稿準備中である。

（３）新たな知見・今後の展望

本研究は、多大な利用価値をもつものの利活用が進んでいない巨大データベースの活用機会を進めることを目的としてきた。そのために、データそのものへのアクセス手段を整備することによってボトルネックを解消するという点に主眼をおいていたが、本研究の過程で、データベース・マネジメントシステム等を利用した効率的な解析技術の有無もまたボトルネックとなるという知見を得た。

ＮＤＢは、当初医療費等の適正化のために活用することを目的としてはじまり、現在では名寄せを容易にできるよう識別子の追加も行われるなど、年々運用方法が改良され利便性・価値の向上が図られてきている。そのため、できるだけ診療エピソードのデータを損なわずプライバシーを保護する手法が求められており、本研究はそのニーズにフィットするものである。しかし、本来のデータベースを用い、エピソード単位での整合性が保持されているか、差分プライバシーによる安全性と有用性にはトレードオフの関係があるがどの程度のノイズがどれほど有用性に影響を与えるか、といったことは、実際のデータを利用し、さらにはドメイン知識を有する専門医の協力なくしては評価できない。そのため今後は、新型コロナウイルスの流行によりやむを得ず変更した本来の研究計画で想定していたように、専門医と連携して実際のデータに基づいた検証・開発研究へ本研究を発展させたいと考えている。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関