Vol.2020-EIP-88 No.15
2020/6/5

医療情報ビッグデータ分析における シャーロックリスクの考察とその形式化

金子 格†

概要:
医療情報処理においてプライバシーの保護は重要な問題である．一般的にある情報の暴露がプライバシーの暴露になるかならないかは，時間によらず安定であるという前提がある．しかし，本報告では，統計情報と推定方法の進歩により個人の属性の推定が，今後著しく変化していく可能性を指摘する．予想しない属性までを知られてしまうリスクを，小説シャーロックホームズに登場する探偵が行う推理になぞらえ，シャーロックリスクと呼び，その例示と形式化を試みる．またシャーロックリスクが高まった場合に，個人の属性情報の暴露による不利益を防止する方法として情報のカプセル化が必要になる可能性を論ずる．

キーワード: シャーロックリスク，機械学習，EHR，プライバシー，匿名化

1. 概要

「そうだね．手を使う仕事をしていた．嗅ぎ煙草を愛用している．フリーメイソンの一員．中国にいたことがある．近頃，大量の書きものをした．これははっきりとわかるが，後はさっぱりわからないよ．」ホームズが笑いつつ首をふるとウイルソン氏は飛び上がった．「ど，どうしてそんなことがわかるんですか！」(一部意訳 金子)．アーサー・コナン・ドイルのシャーロックホームズ，The Red-Hair League[1] の一場面である．

シャーロックホームズは優れた探偵ではあるが，もし仮にこれほどの推定が実際に可能であればプライバシーに対するリスクであるともとらえることができる．

本稿ではこうした高性能な推定による個人情報の暴露のリスクを「シャーロックリスク」と呼ぶことにし，医療情報におけるシャーロックリスクについて考察する．

このようなリスクは一般的には「背景知識攻撃」の一種であるが，本稿ではこうしたリスクが今後より一般的におこることや時間とともに深刻になることをとらえてシャーロックリスクと呼んで以後議論を進める．

またそのようなリスクをコントロールする手段として情報のカプセル化が必要になる可能性を示す．

2. 医療ビッグデータがはらむ課題

2011 の東日本大震災を契機として日本においても EHR(電子化医療情報)の整備が検討され，2020 年から日本版 EHR の二次利用も開始される[2] [3] ．様々な健康情報，投薬情報，将来は遺伝子データベース等のとの連携も技術的には可能と考えられる．一方でこれらの医療情報のプライバシーの管理も重要な課題である．

個人情報保護法では一般的な情報については個人情報の二次利用を許していない．例外として仮名化データは一定の制約の下で，匿名化情報は匿名加工認定事業者による匿名化の元での二次利用が可能となっている．

医療情報の利用においてはこのように公益面からの利用の必要性と個人情報保護という個人の権利の側面のバランスが重要であり一概に強固な保護をすれば済む問題ではなく，一方でセンシティブな情報も含むから特定の個人についてはその個人情報の漏洩がきわめて深刻な権利侵害になり得るという問題をはらんでいる．

ここでよく議論されるのはたとえば「匿名」といった秘匿されるべき属性の秘匿性の安定性である．匿名化についていえば，ある「匿名化」した情報においてその匿名性はどの程度安定しているかという問題である．

本報告ではこのような安定性について議論する．

† 名古屋市立大学
Nagoya City University

そして不安定化をもたらす「推定」がどう行われ，どう定量化でき，どうコントロールされ得るか，という点が本報告の論点である．

3. 匿名化

3.1. 法的定義

まず匿名化について議論していく．個人情報の保護に関する法律(施行日：令和二年一月七日)の第 2 条第 9 項[4] によれば，「匿名加工情報」とは，「次の各号に掲げる個人情報の区分に応じて当該各号に定める措置を講じて特定の個人を識別することができないように個人情報を加工して得られる個人に関する情報であって，当該個人情報を復元することができないようにしたもの」をいう．

また医療情報に関し適用される記述を選ぶとすれば，「特定の個人の身体の一部の特徴を電子計算機の用に供するために変換した」情報がその一つであり，その場合「当該個人識別符号を復元することのできる規則性を有しない方法により他の記述等に置き換える」ことを含め「当該個人情報に含まれる個人識別符号の全部を削除する」といった加工を行うことになる．

また，第 16 条で「あらかじめ本人の同意を得ないで，前条の規定により特定された利用目的の達成に必要な範囲を超えて，個人情報を取り扱ってはならない.」としながら「三 公衆衛生の向上又は児童の健全な育成の推進のために特に必要がある場合であって，本人の同意を得ることが困難であるとき.」は例外としている．

したがって，公衆衛生の向上に役立つとして目的外利用をするという考え方もある．しかしそれは例外的な場合だから，本稿では「個人識別符号を復元できない」つまり個人を特定できないことが要請される場合を想定して以下の議論を進める．

3.2. 匿名化を議論するモデル

匿名化の議論では個人情報を含むデータセットの属性は，は以下のように分類される．

- 識別子
- 準識別子
- センシティブ情報
- その他の属性

これらを簡単に説明する．

- 識別子

識別子は個人識別符号で，識別子を用いれば個人が特定可能な属性である．

- 準識別子

準識別子は単独では個人識別符号ではないが，複数組み合わせることで個人識別符号になりうる属性である．

- センシティブ情報

センシティブ情報は漏洩すべきでない情報，すなわち機微情報である．

- その他の属性

その他の属性とは機微でない情報である．

個人識別符号は登録番号のようなものである必要はない．たとえば個人情報の保護に関する法律についてのガイドライン(通則編)において，DNA のうち個人識別符号に該当するものは様々な塩基配列などの「遺伝型情報により本人を認証できるようにしたもの」とされている．本論では個人識別符号の条件を厳密に論じることは主題ではないがおおむね「個人が認証できる」ことをもって個人識別符号として扱われると考えられる．

つまり匿名化において個人識別符号を削除するためには，残った属性から個人が特定できないようにするべきであると考えられるから，本論ではそのような前提で以下の議論を進める．

3.3. k-匿名化

個人を識別できない，という条件の定義として現在採用されているのが k-匿名化である．k-匿名化とは，与えられた情報から k 人のだれかであるところまでは絞り込めるが k 人のうちのだれかであるかは特定できないことを意味する．

最も低い水準の k-匿名化は 2-匿名化であり，これは 2 人のうちだれであるかが特定できないことを意味する．集団が 2 人であれば 2 人のうちだれであるかわからないということは十分匿名であると言えるが，1 億人中の 2 人と特定されても個人の特定には至っていない．2 匿名化では十分な匿名化とはいえないが k が大きくなれば社会的に許容できる匿名化であると考えられている．

4. 論点

4.1. 匿名化とノイズ

匿名化は元の情報から属性を削除したりノイズを加えたりすることを意味する．識別子，準識別子を削除すれば個人の識別が不可能になることは明らかである．一方で，たとえば身長のような情報でああっても十分精密に測定されていれば個人を識別できる可能性がある．したがって匿名化の手法としてこれにノイズを加えることでその属性から個人を特定できないようにすることが考えられる．

本論ではこのようなノイズの付加の効果を情報理論的に定量化することにする．

たとえば N 人のデータベースにおいて属性 p が

$H_1 = log\ N / log\ 2$ (bit)

の精度で得られていれば情報量 H_1 を持ち，属性 p のみによって個人を特定することが可能である．

その場合 k 匿名化を行うにはおよそ，

$H_2 = log\ N / log\ 2 - log\ k / log\ 2$ (bit)

に精度を低下させるようノイズを加えればよいと考えられる．

4.2. 有益な匿名加工情報は存在するか

このようにノイズを加えて匿名化を行った場合，属性の精度が低下するから有用な情報ではなくなる，という主張がある．そこでノイズを加えても有用な場合があるかを考察する．

今，仮に 20000 人のレコードに体重，血圧，を，ある降圧剤の処方の有無が記録されているとする．個人を特定されないために体重，血圧が均等分布でありその振幅の-20dB のノイズを加えたとすると，体重，血圧によって全体の 1/10 までしか候補が絞られないことになる．したがって体重，血圧の組によってレコードは 1/100 程度に絞り込まれるが 20000 人の 1/100 である 200 人程度にしかしぼりこまれない．

一方体重，血圧の平均値に対するノイズの効果は N 個のランダムノイズは平均化されるため，N の平方根分の 1 となる．降圧剤ありとなしのデータ 1 万人ずつであれ，それぞれの平均値のノイズは-20dB の 1/100 となり，-60dB, 0.1%となる．降圧剤の効果が 0.1%以上あれば効果を検出できる．

このように十分大規模な医療情報データベースでは，ノイズの付加によって k 匿名化行った上で，十分高精度な医学上の統計情報を得ることが可能だといえる．

4.3. 匿名化が棄損される場合

次に匿名化が棄損される場合について例をあげる．

遺伝形質により母集団には 2 形質のグループがありA グループは身長が高いが降圧剤の感度がよく，B グループは身長が低いが降圧剤の感度が低いとする．こうした情報が得られると，同じデータベース，同じレコード属性から絞り込める個人符号は，さらに絞り込める．つまり匿名化が棄損されると予想される．

5. シャーロックリスクの形式化

5.1. シャーロックリスクの定義

以下では本報告シャーロックリスクと呼ぶリスクについて論ずる．まず本報告におけるシャーロックリスクを以下のように定義する．

「すでに提供に合意した本人の情報を利用し本人の予想を超えて個人の属性を推定されてしまうことで本人の不利益(本来ならばそうしたくなかった個人の属性の暴露)を受けるリスク．」

以下では個人情報保護法における匿名性が失われる場合について論じるが，かならずしも、シャーロックリスク＝匿名性の低下，ではない．より一般的にはさまざまな属性の暴露があり得，匿名性の低下はシャーロックリスクの一例である．

5.2. 統計情報による推定の効果

今後，全国 EHR で健康情報が蓄積された場合に，医療統計情報が増加することはもともと期待されることであり，抑制すべきではない．しかしそのような統計情報の増加によって，匿名性が低下する可能性があることを以下で示していく．つまり，日々拡大する統計情報と組み合わせることにより k 匿名性の安定性が失われる可能性について考えてみる．

外部の統計情報により属性から個人の識別が容易になる例を図 1 と図 2 を使って説明する．

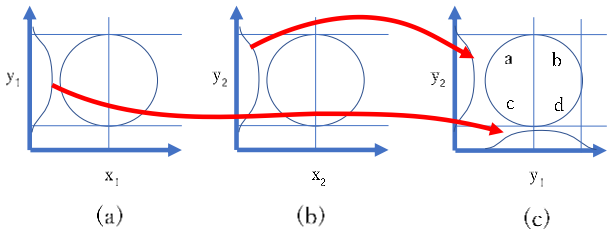

図 1 統計情報の取得前

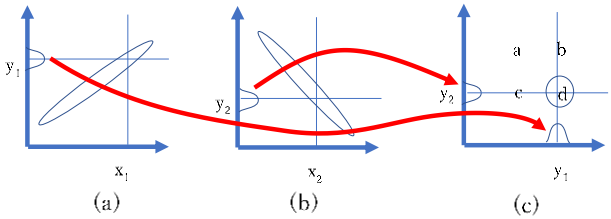

図 2 統計情報の取得後

図 1(a)と(b)は属性 x_1，属性 x_2 が与えられた場合に属性 y_1，属性 y_2 がどう推定できるかを表している．(c)は y_1, y_2 が与えられた場合にどう個人が特定できるかをあらわす．

x_1 と y_1 にも，x_2 と y_2 にも相関がないので x_1,x_2 が与えられても y_1, y_2 の分布は広い．したがって y_1, y_2 から個人を特定することができない．

図 2 では，統計情報の集積により x_1 と y_1 にも，x_2 と y_2 にも高い相関が得られた場合を示している．(a)と(b)は属性 x_1，属性 x_2 が与えられた場合

に属性 y_1，属性 y_2 を推定するが，統計情報が集積されたことで y_1, y_2 の範囲は狭まる．
結果として(c)で y_1, y_2 から個人が特定される．

まとめると，属性 x_1, x_2 によって図 1 の状況では個人識別ができないが，図 2 の状況では個人識別が可能になる．そして図 1 と図 2 の差異を生じさせているのは x_1 から y_1, x_2 から y_2 という統計情報である．つまり統計情報の強化により，既存の属性からこれまでよりも高精度な情報が引き出せるようになることがあり，その結果，属性情報の匿名性が低下する可能性がある．

5.3. 情報量に着目した考察

この状況を情報量に着目すると，どうモデル化されるだろうか．

図 2 における(c)のテーブルは属性 y_1, y_2 から個人の特定が可能な状居をあらわしている．この場合個人の特定にエントロピー H が必要だとすると，結局 H 相当の情報量が個人情報と結びつけられた属性情報に関して得られれば個人識別が可能である，ということになる．

式[1]に情報量の関係を示す．$x_1 \cdots x_L$ が属性情報の元となる情報源，$n_1 \cdots n_M$ はノイズ，Bv 情報源とノイズの混合方法，$y_1 \cdots y_K$ はそれに A を乗じて属性情報の推定を行うことを表している．B が未知であると Bv の見かけ上の情報量は少ない．その場合いかなる A を用いて y を求めても y の情報量は Bv の見かけの情報量を超えることはない．もし Bv のみかけの情報量が十分小さければ，匿名化は十分である．

しかし B が精密に推定できればその知識を使って Bv に含まれるノイズをキャンセルすることが可能である．そのようなノイズキャンセリングを行うと Bv の情報量が増加する．それによって y の情報量が増加して匿名性が低下または劣化する．

$$\begin{pmatrix} y_1 \\ \vdots \\ y_K \end{pmatrix} = ABv = AB \begin{pmatrix} x_1 \\ \vdots \\ x_L \\ n_1 \\ \vdots \\ n_M \end{pmatrix} \quad [1]$$

したがって，このモデルではシャーロックリスクは B の統計情報を蓄積することにより y から得ら得る最大の情報量が変化することである．このようなモデルを用いれば必要なノイズレベルとくらべ現在どの程度の「余裕度」があるかを定量化できる可能性がある．そしてリスクを定量的に評価したり制御することが可能になると期待できる．

6. ビッグデータによる影響

6.1. 医療情報のビッグデータ化

医療の世界では情報のビッグデータ化が世界的な潮流となっており，日本においても 2020 年から日本版 EHR の二次利用も開始される[2][3]．医療情報のビッグデータは製薬や医療の発展に大きな効用があると考えられている．さらに遺伝子解析の分野では 2007 年以後劇的なコストダウンが進み，ムーアの法則を超える速度で低コスト化と普及が進んでいる[5]．2000 年に世界的プロジェクトによって達成された人間の全遺伝子配列スキャンも現在ではたった 10 万円台で行うことができる．

これら収集される情報は様々な医療的研究に利用でき今後健康情報，投薬情報，将来的には遺伝子データベース等の連携は飛躍的に進むだろうと予想される．

6.2. 分析能力の向上

一方こうしたビッグデータの分析手法の進歩も著しい．近年機械学習やデータサイエンスの発達が著しく，今後も分析方法の向上は続くと予想される．また，データベースの大規模化によっても学習データの増大により分析精度が向上すると予想される．

その結果医療情報を利用した様々な属性の推定は劇的に進む可能性がある．そうして得られた情報の扱いをどう考えるべきか，が当然提起されるだろう．

7. シャーロックリスクの問題点

7.1. データ取得時の同意

このような状況においてシャーロックリスクは何が問題だろうか．個人情報の扱いは個人情報保護に関する諸制度に適合する形で行わなければならいことはいうまでもない．したがって医療情報の収集においても同意を得た患者などから許可を得てデータを取得し，了解は十分とれているということが前提となる．

データの利用，分析結果の公表は，いずれも提供者の了解のもとにデータを取得，分析しているので，その点は問題がないと考えられている．しかしデータから得られる個人の属性情報が時間とともに変わってくるとするとこの同意が時間とともに本人の意思と乖離していくと考えらえっる．この点に充分考慮しないことは形式的には同意原則に従っていてもデータ提供者の不審や不満を招く恐れがある．

7.2. 入力データ提供の必要性

こうした分析において分析の元となるデータを公表する必要は必ずしもないという考え方もある．しかし研究結果の公平性や信頼性の確保という観点からかんがえれば，オープンサイエンスが現在の潮流であり，またデータの提供者からみた信頼性の確保のためにも，分析前のデータの提供や公表の必要性は高いと考えられる．したがって，少なくとも研究目的で同意を得て実験を行う場合，入力データと分析結果は揃えて提供，公表されることが望ましいと考えられる．

7.3. 分析結果における問題の可能性

その場合，当然データ提供者はできれば必要以上に自分の属性を公開したいとは思えないから一定範囲の情報を公開するつもりでデータを提供する．しかしシャーロックリスクの存在は，分析方法の改良によって得られた分析結果が，当初の予想を超えた範囲に及ぶ可能性を示している．そこでそのような当初の予想を超えた属性の公表をどうコントロールするか，という課題が生じる．

たとえば，身体活動以外の分野では，音声の分析によりガンの発生を検知する可能性があることが示されている．また公開されている SNS の発言の内容から様々なプロファイリングが可能である．

身体活動の分析においても，分析可能な結果の中に，提供すべきであるか，公開すべきであるか，議論が必要なものが生じる可能性があると言える．

7.4. 個人情報保護法での限界

個人情報保護法では同意原則があるのでそれによって本人の同意はとれているから問題がない，という考え方があるがシャーロックリスクの存在はその原則では不十分である可能性がある．つまり冒頭の例のように同意した情報の公開によって，「いったいどうしてそんなことが分かったのですか？」という事態が生じる可能性があるからである．

7.5. 分析の制限の可能性

ではすでにデータを提供してしまったものが，予想をこえた個人の属性の暴露による不利益を受けることから守るために，どういう方法がとり得るだろうか?

まず分析手段を所有するものが結果の提供を控える，ということが考えられる．しかし分析方法が入手できれば，分析はだれでも行うことができる．次に，分析方法の提供を控えることが考えられる．しかし分析方法は様々な人が改良可能であるので，この方法で制約してもやがては他のルートで分析方法が可能になり分析結果もアクセスできることになる．

次に，分析する行為や分析方法の所持を制限するという考え方もある．著作権侵害の防止のためにコピー保護の迂回手段の提供や保持が規制されているがその考え方に近い．

しかしデータの分析手段は単なるアルゴリズムにすぎない．コピー保護の迂回手段が一般になんらかの特許や特殊な仕様を持っているためにコピー保護の迂回手段であるという特定がしやすく，規制がしやすいが，ビッグデータの分析手段は汎用的なものなのでアルゴリズムが一旦発見されればだれでも実装利用することができる．制限をまもらずに分析方法を保有，利用しても，外部からは検証することはきわめて困難である．

7.6. 提供した情報の分析を制限する合理性

データの加工によって元データからは従来得られなかった多くの情報が得られるようになった場合に，その公表を制約することは有益か，という点を検討してみる．

情報を加工提供する行為に対する制約について，利用者にとって不利益となる分析結果についてどのような制約が可能であるかを考えてみる．

この場合，提供された情報の分析結果の公表に制約を課すれば効果的な防止手段になるだろうか．分析の元データは提供可能であるとすると，情報提供を受ける側は，元データの移転は妨げられていないから，元データを入手できると考えられる．したがって分析結果を移転しなくても分析方法が知られれば，すでに得た元データと分析方法を用いて，新たに推測可能となった情報がアクセスできる．つまり分析結果の公表や移転を制約しても，けっきょくそのような分析結果が取得されてしまうから，そのような規制にはあまり意味がないと思われる．

しかし，民法の名誉棄損は，たとえ事実であってもそれを公然と社会的評価低下させる恐れがあれば成立する．個人情報保護法も入手方法と関係なく「個人と紐づいた情報」を規制している．これらは，仮にそうした情報を入手する方法があったとしても情報を拡散することを防ぐことを意図しているから，仮に分析手法を使う事でだれでもその情報を得られるとしても，分析を行ったり，結果を記録，交換したり，公表や移転したりすることを制約することに意味がある，という主張も可能と思われる．

7.7. シャーロック問題とその検討

データの集積と分析の高度化がはらむ問題の

本質は，分析技術の高度化によりどの程度の分析が可能であるかということの不確実性に起因する．この不確実性は，分析技術の急速な高度化とデータ集積の大規模化が背景になっている．

より大規模なデータベースと対照することにより多くの特徴と母集団の比較をより精密に行うことが可能となる．また，その統計的な処理方法も急速に進歩している．結果としてこれまでにくらべ，少ない情報から多くの推定を引き出せることが可能になりつつある．

今後，データベースの大規模化と分析技術の進歩が続くと考えると，それらの情報により予想外の推定が可能となる，という状況が予想される．どのような可能性があるかという認識を共有し，そのような状況にどう対応するかを検討する必要があるだろう．

7.8. カプセル化されたデータの分析

一つの解決方法としてはデータをカプセル化する方法である．図 3 は通常のデータ処理のデータフローであるがデータ提供者はデータ提供ジにどのような分析手段を利用しだれがどのように分析結果を利用しているかの許諾を得るが実際にどう分析されだれが利用しているかという情報を得る手段は処理システム上では提供されていない．

提案するのは図 4 に示すカプセル化されたデータ処理である．この場合コントロールやデータの流れはすべて双方向の制御を実装している．したがってデータ提供者は分析手段や分析結果の提供を事前に許諾する自由があるだけでなく利用が続いている間常時それらを監視したり許諾を取り消したりすることができる．

図 3 通常の分析のデータフロー

図 4 カプセル化されたデータの分析

このようなデータ提供手段はたとえばマイナンバーの一時番号の利用ではすでに実現している．マイナンバーと一時番号を結合する操作を行えばマイナポータルでマイナンバー保持者はそれを知ることができる．また音楽のストリーミングサービスも同様のモデルでサービスを行っている．サービス利用者はデータを永久的に取得するのではなく「利用」にたいして許諾を得て対価を支払う．

同様の機構を個人の属性情報の利用について行えば，事後的に予想以上の個人の属性情報の推定が行えるようになった場合にある程度対応が可能になると考えられる．

8. まとめ

本報告では，匿名化された個人情報の匿名性が統計情報の増加により低下する減少について考察した．またそのような形で匿名性が低下した場合にも個人情報を保護するには，「分析に対する制約」を導入する必要があるかもしれないということを考察の結果として示した．

個人情報保護に関する考察では，当初予想していなかった身体活動の推定が可能となる可能性があることを指摘し，実験結果の公表において入力となる分析前のデータを公表する場合における問題点を考察した．

最後にシャーロックリスクの分析や対処方法を考察し，解決手段の一つとしてデータのカプセル化を提案した．

今後大規模な医療情報ビッグデータの構築やその分析手法が急速に発達すると考えらる．当初予想した以上の分析が可能になるという状況が今後は発生すると予想している．こうしたリスクをよく理解しそのような状況にあった制度設計の議論を進めていくことが必要ではないかと考える．

References:

[1] Sir Arthur Conan Doyle, The Red-Headed League, The Strand Magazine, August, (1891)
[2] 吉原 博幸，千年カルテプロジェクト：本格的日本版 EHR と医療データの 2 次利用に向けて，情報管理，60 巻 11 号(2017)
[3] 田中 博，日本版 EHR（Electronic Health Record）の実現に向けて，情報管理，第 54 巻 9 号(2011)
[4] 個人情報の保護に関する法律，https://elaws.e-gov.go.jp/search/elawsSearch/elaws_search/lsg0500/detail?lawId=415AC0000000057#5, 2020/4/23 取得
[5] Kris A. Wetterstrand, M.S., The Cost of Sequencing a Human Genome, NIH National Human Genom Research Institute, https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost, 2020/4/23 取得