連結可能匿名化のための暗号手法

岡本（オカモト）　悦司（エツジ）*

レセプトとカルテといった異なったデータを個人単位で連結（リンケージ）したり，複数の機関から個人情報を収集するがん登録のような疾病登録事業を，個人名を平文のままで行うのではなく暗号により連結可能匿名化して行う可能性を検討する。

暗号化と解読の両方が必要な通信とは異なり，暗号化のみで足りる研究目的の連結（リンケージ）の場合，情報提供を受ける研究者が鍵を共有する必要はないので自治体や保険組合等のデータ保有者は簡単な暗号化により安全に研究者にデータ提供を行うことが可能である。

Microsoft エクセル®を用いた人名暗号化の具体的な手法を紹介する。人名の漢字を JIS コード化し，そのコードを無作為に選んだアルファベット（大小52文字）で置換する。この数字とアルファベットの対応表が鍵であり，5.74×10^{16} 通りの組み合わせがあることから鍵無しに解読は不可能である。これにより万一漏洩があってもプライバシー侵害が起こらない技術的担保ができ，公衆衛生研究が促進されよう。

がん登録や脳卒中登録のような複数の機関から個人情報を収集し追跡する疾病登録事業においても公開鍵暗号を用いることにより連結可能匿名化された登録システムが可能になる。しかしながら，暗号化作業が複雑であること，登録機関からの問合わせが不可能であること，鍵を公開することにより人名と暗号との対応表を誰でも作れることから安全性は十分には保証されず，暗号のみに頼って疾病登録事業の連結可能匿名化はなおも困難である。

Key words：暗号，個人情報保護，レコードリンケージ，がん登録，連結可能匿名化，疫学研究倫理指針

Ⅰ　緒　　言

人を対象にした公衆衛生研究を実施するうえで，個人情報の扱いはきわめて重要な問題であるが，2002年7月に「疫学研究に関する倫理指針（以下，疫学指針）」が，2003年5月には個人情報保護法が施行され，学術研究においても個人情報保護の必要性はいっそう高まっている。

個人情報とは「生存する個人に関する情報であって，当該情報に含まれる氏名，生年月日その他の記述等により特定の個人を識別することができるもの（他の情報と容易に照合することができ，それにより特定の個人を識別することができることとなるものを含む。）」（個人情報保護法第2条）と定義される。すなわち個人から得られた情報であっても誰のものか特定できない連結不可能匿名化された情報は個人情報には該当しない。疫学指針も連結不可能匿名化された情報のみを用いる研究は対象にしていない（疫学指針第12②）。

しかしながら匿名化されていても連結可能（他の情報と容易に照合することができ，それにより特定の個人を識別することができること）な個人情報については疫学指針が適用され，倫理審査等の対象とされる。当然ながらそこでは個人情報保護の具体的な扱いが重視される。

疫学指針は疫学以外の公衆衛生研究（たとえば心理，経済研究）には適用されず[1]，個人情報保護法も憲法で保証された学問の自由を侵害しないよう「大学その他の学術研究を目的とする機関若しくは団体またはそれらに属する者」は学術研究の用に供する目的の範囲内において個人情報取扱事業者の義務が除外される（同第50条）。しかしながら「個人データの安全管理のために必要かつ

* 国立保健医療科学院
連絡先：〒531-0197 埼玉県和光市南2-3-6
国立保健医療科学院研究情報センター　岡本悦司

適切な措置，個人情報の取扱いに関する苦情の処理その他の個人情報の適正な取扱いを確保するために必要な措置を自ら講じ，かつ，当該措置の内容を公表する」努力義務は課せられる。

連結不可能匿名化の作業は氏名や住所等をデータから削除すればよいだけだから容易であり，セキュリティ上の問題も少ない。レセプト（診療報酬明細書）を例にとると，たとえば1,000件のデータを傷病別，男・女別あるいは医療機関別に日数や点数を集計することは通常の業務統計であってインフォームドコンセントや倫理審査も必要ない[2]。

しかしながらレセプトを，たとえば検診カルテのような他の個人データと連結（リンケージ）するとなると事情は異なる。1,000件のレセプトからインフルエンザの日数や点数を算出することは通常の業務統計だが，1,000件のレセプトをワクチン接種者と非接種者に分けて日数点数を比較しようとすると連結不可能匿名化のままでは行えない[3]。そのためには，たとえば保健センターが保有するワクチン接種者台帳とレセプトとを連結しなければならない。同様のことは，アンケート調査票，検診カルテそしてレセプト等をリンケージするコホート研究においても起こる[4]。

総背番号の無いわが国では，レコードリンケージは専ら氏名や生年月日，住所によって行われる。総務省の調べによると2002年4月現在で都道府県および市町村の65.7％が個人情報保護条例を有しており，氏名等を自治体外の研究者に提供することはたとえ研究目的であっても個人情報保護審査会の承認等制約が加えられることが多い。各種保険組合についても個人情報保護の徹底を促す通知も出されている[5]。今後個人情報を使った研究を実施するにあたって，個人情報保護担当者の理解と納得を得ることが重要となる。

II　暗号化の必要性

個人情報を提供する場合，個人情報保護の担当者が最大の懸念を示すのが万一の漏洩である。こうしたリスクに対して「安全管理を徹底させる」「守秘義務を明記する」といった漠然とした説明では不十分であり，ましてや研究そのものの意義や重要性を力説しても説得力に乏しい。万一漏洩しても決してプライバシー侵害は生じない技術的担保を示すことができなければ，いかに有意義で重要な研究も行えなくなるおそれがある。

どんなに扱いを厳重にしても機密情報は漏れることを古来より人類は知っていた[6]。漏洩することを前提に，漏洩して第三者に情報が渡っても内容をわからせないようにする技術が暗号化である。暗号化が完全であれば，暗号そのものはもはや秘密にせず公開さえしてもよくなる。

米国CDCは全国調査の個票データをインターネット上で公開している。そのうちのひとつNAMCS（National Ambulatory Medical Care Survey，全国外来調査。わが国の患者調査に相当する）2001年版は2万4281件の外来個票データからなり，最大3の診断名，最大6の薬剤名などに加えて医師も匿名化されて含まれている。ちなみに著者がCDCサイト（ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NAMCS/）よりダウンロードして分析したところ，1,013人の医師が4ケタの番号で表示されており，この番号で連結し個票件数を集計したところ最も件数が多かったのはNo. 629の医師で112件あった。

わが国ではこうした個票の使用は統計法等により厳格な手続きが求められるが，効果的な暗号技術を活用することによって個票がより使いやすい環境整備が望まれる。

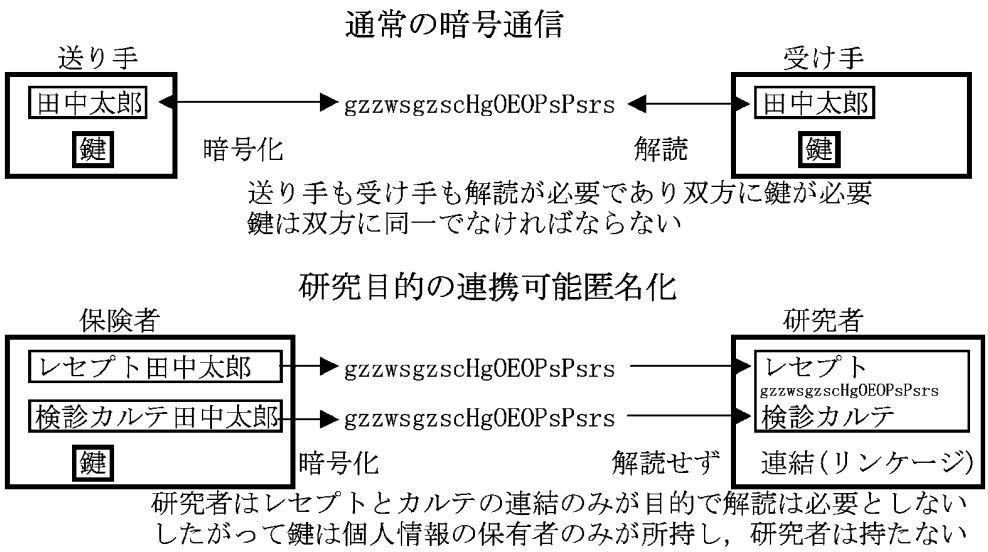
III　リンケージと通信の相違

暗号は通信手段として開発された。送り手と受け手がおり，送り手は文章を「鍵」と呼ばれる変換表にしたがって暗号化して送信する。受け手は受け取った暗号を同じ鍵を使って元の文章に解読（復号化）する。暗号が途中で盗聴者に盗まれても鍵がないので元の文章に戻すことはできない（図1）。受け手は確実に元の文章に戻せなければならないが，それ以外の者は元の文章に戻すことができてはならない。すなわち暗号化と解読が同時に可能でなければならない。

そのためには予め，双方が同一の鍵を共有しておかなければならないが，その鍵をどうやって相手に安全に届けるか，が問題になる。鍵は最重要機密だが，その鍵を安全に届けられる手段がもしあるなら暗号化などそもそも不必要だからである。

このいわゆる「鍵共有問題」は1977年に公開鍵（RSA暗号）が開発されるまで暗号学における最

図1



大の課題であった[7]。また同一鍵を共有しなければ機密情報をやりとりすることはできず，後述する疾病登録のように不特定多数の医療機関から個人情報を提供してもらうことには適さない。何より同一の鍵を多数に共有させること自体漏洩の危険が高まる。

ここで，通信とリンケージの違いを検討する。リンケージでは，レセプトとカルテを連結する研究者は人名を解読する必要はなく，むしろ解読できない方がよい。すなわち研究者はgzzwsgzscHgOEOPsPsrsと記載されたレセプトとカルテが同一人名であることさえわかればよく，それが田中太郎を意味することを解読する必要もないしまたしてはならない。その点で研究者も盗聴者も同一である。

通信では暗号化と解読の両方が必要になる双方向だが，リンケージでは暗号化のみの一方通行でいい。したがって鍵を所有する必要はなく鍵共有問題も起こらない。このことはリンケージ手段として暗号を用いる場合にきわめて有利な条件であり，公開鍵のような高度な技術を用いなくても，万一漏洩してもプライバシー被害の生じないリンケージが可能であることを示す。

Ⅳ　エクセルによる簡便な人名暗号化

広く普及している表計算ソフトエクセルでできる簡便な人名暗号化手法を提案する。

基本的な原理は人名の漢字をJISコードに変換し，それぞれの数字を乱数発生させたアルファベットに置換する，という単純なものである。原理は単純だが，アルファベットの大文字小文字を組み合わせれば5.74×10^{16}通り（＝PERMUT（52, 10））の組み合わせがあり，暗号だけから人名を解読することはほぼ不可能である。

1.　文字のコード化（encoding）

まず漢字のコード（数値）化を行う。すべての漢字はJISコードがあり，エクセル上ではcode関数で変換できる。田中太郎の場合[8]

＝code（“田”）-〉17732
＝code（“中”）-〉17254
＝code（“太”）-〉16960
＝code（“郎”）-〉20282

となる。このコード化だけでも匿名化にはなるが，これらの数字が人名であり，JISコードであることは容易に推測でき，推測できれば容易に解読できるので暗号化としては不十分である。

2.　鍵の作製

0〜9の数字をアルファベットに変換する「鍵」を作製する。任意に決めてもよいが，任意に決めると個人のクセ（覚えやすいアルファベットにしてしまう等)がでるので，乱数を用いた方がよい。52文字（大小文字）の横の列にrand（　）関数

を挿入して乱数を発生させ，その乱数でソートして上位10文字を選択する。たとえば以下のような組み合わせが鍵になる。

0 1 2 3 4 5 6 7 8 9

P g s w H c O z r E

鍵はきわめて重要であり，暗号化したデータ保有者（自治体，保険組合等）が厳重に管理し，研究者はもちろん外部に一切漏洩してはならない。

3. 鍵による文字変換

エクセルメニューより編集->置換を選択し「大文字と小文字を区別する」をクリックしたうえで「検索する文字列」に0,「置換する文字列」にPを入力し「全て置換」する。これを鍵にしたがって0から9まで10回くりかえす。その結果，田中太郎は以下のように暗号化される。

＝code（“田”）->17732->gzzws

＝code（“中”）->17254->gzscH

＝code（“太”）->16960->gOEOP

＝code（“郎”）->20282->sPsrs

4. 暗号化の完成

田中太郎という人名はgzzwsgzscHgOEOPsPsrsと暗号化された。

この暗号は5文字ずつ区切って鍵と逆の置換を10回くり返せば元のJISコードコードに戻り，さらにchar関数を用いれば元の漢字に戻る。

gzzws->　＝char（17732）->田

gzscH->　＝char（17254）->中

gOEOP->　＝char（16960）->太

sPsrs->　＝char（20282）->郎

しかしながら鍵を持たない誰か(研究者も含め)が，この暗号だけから人名を解読することは不可能であり，よって暗号が第三者に読み取られてもプライバシー被害は生じない。

V　疾病登録への公開鍵暗号の応用

がんや脳卒中といった疾患の患者を登録し，複数の医療機関ひいては死亡個票と連結して追跡調査を行うのが疾病登録である。疾病登録は同一患者の追跡という目的から匿名化できず，氏名や住所といった個人情報を収集している。疫学指針では，他機関から匿名化されていない既存資料の提供を受けるにはインフォームドコンセントを原則としている（指針11⑵②）が，がん登録は保健事業であるとして指針対象外とされている[9]。

しかしながら同意はおろか本人に告知されないまま保健事業に参加させられるなど不合理である。疫学指針は匿名化されておれば連結可能であってもインフォームドコンセントを略して資料の提供を認めている（指針11⑵①）。そこで暗号により連結可能匿名化されたかたちで疾病登録機関が患者情報を収集する可能性を検討する。

上述の単純な暗号化では，情報の送り手はすべて鍵を有しなければならない。疾病登録では全ての医療機関が送り手，中央の疾病登録機関が受け手になる。したがって，すべての医療機関に共通の鍵を予め配布しなければならない。

共通鍵で暗号化するから田中太郎はどの医療機関でもgzzwsgzscHgOEOPsPsrsと暗号化されるのであって，異なった鍵を使うと異なった暗号になり疾病登録機関は追跡不能になってしまうからである。また鍵の配布を受けた医療機関は他の医療機関の患者暗号を自由に解読できるし，多数の医療機関に最重要の鍵を配布すること自体，漏洩が不可避になり，この方法は使えない。

つぎに公開鍵暗号を検討する。公開鍵とは開発者のイニシャルをとってRSA暗号とも呼ばれ，インターネットで広く使われている。その原理の詳細は他書[10]に譲るが，かいつまんで説明すると以下の通りである。

1. 公開鍵

従来の暗号では送り手と受け手が共通鍵を有し，その鍵は暗号化も解読も両方できる。よって共通鍵は送り手受け手双方にとって絶対秘密としなければならない。

ここでA鍵，B鍵の2つを考える。ABはいずれも暗号化か解読かいずれかしかできず，相互に鍵と鍵穴の関係にある。すなわちA鍵で暗号化された暗号はB鍵によってのみ解読でき，B鍵で暗号化された暗号はA鍵によってしか解読できない。A鍵は自分で暗号化した暗号を自分で解読できず，それはBについても同様である。

このうちA鍵は公開し，反対にB鍵は秘密にして受け手が保有する。受け手は「自分に情報を送る時は公開されたA鍵を使って暗号化して送れ」と意思表示する。このようにして送られた暗号を公開されたA鍵を有する盗聴者が取得してもA鍵では解読できないので安全である。このA鍵を公開鍵と呼び，こうした仕組みにより不特

定多数との暗号やりとりを可能にするのが公開鍵暗号である。

疾病登録では，多数の医療機関が送り手，一つの疾病登録機関が受け手である。匿名化して疾病登録を行うとしたらA鍵のみ公開し，B鍵は作らないようにすればよい。これにより疫学指針に合致した連結可能匿名化されたかたちでの疾病登録が可能になる。

2. 疾病登録への応用の限界

このように公開鍵暗号による匿名化された患者情報の収集は可能であるが，実現の可能性となると現時点ではまだ困難といわざるをえない。その第一の理由は，高度な技術とコンピューター機能（数千ケタの演算能力が必要となる）が要求され，素人がエクセルでできる程度のものではないこと。それを多数の医療機関に求めることは無理である。

第二の理由は，疾病登録機関からの問合わせができない，ということである。がん登録では，住所地の自治体に問合わせたり住民票を閲覧したり死亡個票と照合する等によって登録患者の生死を調査することがあるが，個人名がわからないのではそのような調査はできなくなる。

そして第三の理由は，人名という個人情報の脆弱性である。鍵が公開されるので暗号化は誰でも行える。しかも人名は有限なので，人名を片っ端から暗号化し，取得した暗号と照合すればどの暗号がどの人名なのか解読される可能性が高い。現実に使われているRSA暗号では，たとえば電子投票などで人名のような脆弱性の高い情報を暗号化する場合には乱数を加えること（パディング）で対応している。パディングを行っても解読に支障はない。ところが，疾病登録に暗号化を導入する目的は，解読を行わず（解読するのであれば氏名をそのまま収集している現状と変わりなくなる）に暗号のみで患者を追跡することであった。もしセキュリティを考えて乱数を加えるともはや暗号だけでは患者を追跡することが不可能になる。

このように公開鍵暗号を疾病登録に応用するには限界があるが，やりとりされる暗号そのものの管理を厳重にし，漏洩を防止すれば少なくとも匿名化して収集するという目的は達せられよう。

Ⅵ 結　　語

人名を暗号化することにより個人情報保護をはかりつつレコードリンケージを行う具体的方法を提案した。暗号化は素人でもエクセルで容易に行うことができ，リンケージだけが目的で解読を行わないのであれば，漏洩してもプライバシー被害につながらない技術的理論的担保を与えることができる。

多数の医療機関を対象にした疾病登録でも，高度技術を要する公開鍵暗号によって連結可能匿名化した患者情報の収集は可能である。ただ，その暗号化には脆弱性が伴い，それに対する乱数を付加するといった対策も「暗号化のみで解読はしない」リンケージの有利性がかえって足かせになって不可能である。脆弱性は暗号そのものの管理を厳格に行うことによってある程度解決できるが，登録機関が個々の患者について問合わせ等を行うことができなくなる，といった制約のため暗号化のみに頼った疾病登録の連結可能匿名化は困難である。

人名の暗号化は，むろん，個人情報保護の面で完全ではない。同姓同名者の区別は不可能なので，より正確に個人を区別するためには生年月日も加えて暗号化する等より複雑な処理が必要になる。またたとえ人名は解読できなくても，年齢や傷病名，また診療月といった他の情報により個人を特定できる可能性はつきまとう。

それでもなお，簡便かつ安全な暗号化によって，漏洩によるプライバシー被害を起こさせない技術的担保を与えられることは，レセプト，検診カルテ，アンケートといった情報を有機的につなぐレコードリンケージ研究を促進する効果が期待できる。

本研究は厚生労働科学政策科学推進研究事業「レセプト情報の利活用と個人情報保護のあり方に関する研究（H14-政策-016。主任研究者：小林廉毅）」の一環として行われた。

（受付　2003.11. 3
採用　2004. 3.18）

文　　献

1） 岡本悦司．公衆衛生研究における「疫学研究に関する倫理指針」の適用．日本公衛誌 2003; 50:

1079–1090.
2） 岡本悦司．レセプトの法的性質と研究利用の可能性．日本公衛誌 1995; 42: 999–1006.
3） 岡本悦司．高齢者インフルエンザ予防接種の効果研究．小林廉毅，編．平成14年度厚生労働科学研究費補助金政策科学推進研究事業報告書「レセプト情報の利活用と個人情報保護のあり方に関する研究」．東京：東京大学大学院医学系研究会公衆衛生学．2003; 66–77.
4） 岡本悦司．コホート研究における医療費レセプトリンケージ―法的，制度的，技術的側面から．岸玲子，編．平成11年度厚生科学研究費補助金（長寿科学総合研究事業）報告書「高齢期における活動的生活維持のためのサポートネットワークの役割に関する研究」．札幌：北海道大学医学部予防医学講座公衆衛生学分野．2000; 93–131.
5） 厚生労働省保険局国民健康保険課長．保険者における個人情報保護の徹底について．保国発第0314001号［平成15年3月14日］
6） サイモン・シン（青木 薫訳）．暗号解読（原題 ‘‘The Code Book–The Science of Secrecy from Ancient Egypt to Quantum Cryptography’’）．東京：新潮社，2001; 17–197.
7） 伊藤正史．暗号理論．ナツメ社 2003: p. 90.
8） Code関数は1文字ずつしかコード化できない．人名のように連続した文字を多数一度に切りわけるにはmid関数が便利である．田中太郎という文字列の3番目から1文字のみぬきだすには＝mid（“田中太郎”，3, 1）-〉太となる．
9） 施行通知（平成14年6月17日）別添3.「疫学研究に関する倫理指針」とがん登録事業の取扱いについて．
10） 結城 浩．暗号技術入門．東京：ソフトバンク，2003; 106–143.

ENCRYPTION TECHNIQUE FOR LINKABLE ANONYMIZING

Etsuji OKAMOTO*

Key words：encryption, privacy protection, record linkage, cancer registry, unlinkable anonymizing, the Ethical Guildelines for Epidemiological Research

Linkage of different records such as health insurance claims or medical records for the purpose of cohort studies or cancer registration usually requires matching with personal names and other personally identifiable data. The present study was conducted to examine the possibility of performing such privacy-sensitive procedures in a "linkable anonymizing" manner using encryption.

While bidirectional communication entails encryption and deciphering, necessitating both senders and receivers sharing a common secret "key", record linkage entails only encryption and not deciphering because researchers do not need to know the identity of the linked person. This unidirectional nature relieves researchers from the historical problem of "key sharing" and enables data holders such as municipal governments and insurers to encrypt personal names in a relatively easy manner.

The author demonstrates an encryption technique using readily available spread-sheet software, Microsoft Excel® in a step-by-step fashion. Encoding Chinese characters into the numeric JIS codes and replacing the codes with a randomly assigned case-sensitive alphabet, all names of Japanese nationals will be encrypted into gibberish strings of alphabet, which can not be deciphered without the secret key. Data holders are able to release personal data without sacrificing privacy, even when accidental leakage occurs and researchers are still able to link records of the same name because encrypted texts, although gibberish, are unique to each name. Such a technical assurance of privacy protection is expected to satisfy the Privacy Protection Act or the Ethical Guidelines for Epidemiological Research and enhance public health research.

Traditional encryption techniques, however, cannot be applied to cancer or stroke registration, because the registrar receives reports from numerous unspecified senders. The new public key encryption technique will enable disease registry in a linkable anonymizing manner. However various technical problems such as complexity, difficulties in registrar inquiries and risk of code-breaking make the encryption technique unsuitable for disease registry in the foreseeable future. (320 words)

* National Institute of Public Health