同意未取得の医療情報利活用に向けた匿名化技術の適用可能性検証

栗原　幸男
高知大学医学部看護学科 保健医療情報学教室　教授

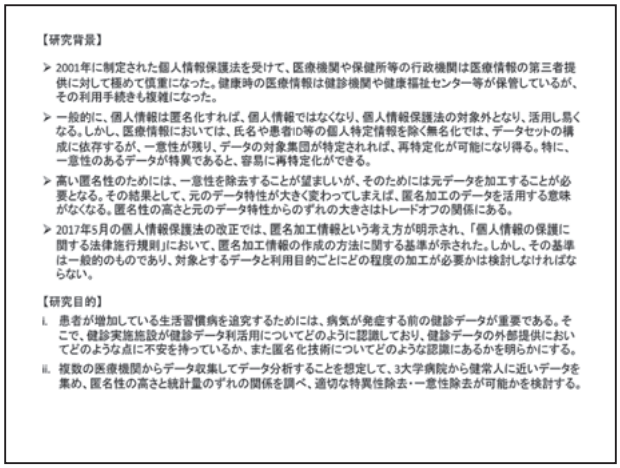
【ポスター1, 2】

皆さんご存じのように、医療における情報化が進んでいて、電子化されたデータがどんどん増えてきています。それを有効に使っていくのは、今後の医療を進めていく上での大事な課題だと思います。

今年の5月の終わりに個人情報保護法が変わり、若干、匿名化したデータの扱いに関してどうするかということが出ましたけれども、この研究を始めた時点ではまだそこは出ていなかったので、私自身、病院のデータを利用させていただいて、いろいろな研究をさせていただいています。その中で、自分自身の経験として、例えば、地域との比較をしようということで保健所からデータを得ようとすると、いろいろとハードルがありました。ケースによっては、市議会で審議していただいたこともあります。

そういう中で、個人情報をどう扱っていくかということに関しては、基本的には匿名化をして個人を特定できない状態になれば、一応利用できるということになっています。今までは割と簡単なやり方で、無名化という言葉を使いましたが、要するに、個人の、名前とかID番号とかを除く、そういうところだけでOKということをやっていました。しかし、今回の個人情報保護法の改定では、より厳しく、個人を特定できないような形で匿名化し、再特定化ということができないようにする。基本的な情報を除いたとしても、いろいろな特殊性で…よく例に出されるのは、小学生で180cmとか170cmの子どもというのは特異なことなので、いろいろな情報を除いたとしても個人が特定されてしまう。そういう特異的な情報の扱いをしっかりやらないといけないということで、どんなふうにしたらいいか。

ポスター1


【研究背景】

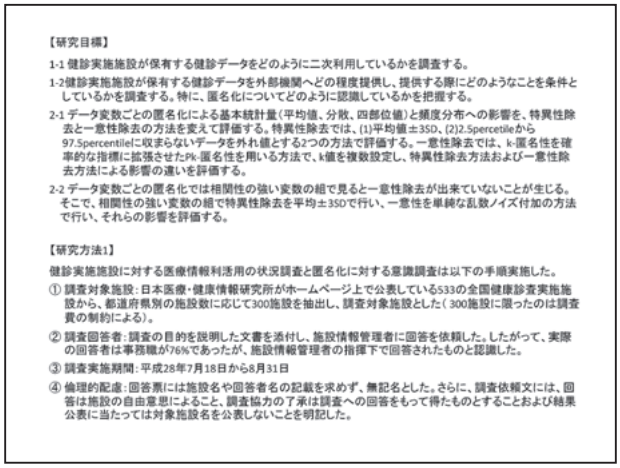
- 2001年に制定された個人情報保護法を受けて、医療機関や保健所等の行政機関は医療情報の第三者提供に対して極めて慎重になった。健康時の医療情報は健診機関や健康福祉センター等が保管しているが、その利用手続きも複雑になった。
- 一般的に、個人情報は匿名化すれば、個人情報ではなくなり、個人情報保護法の対象外となり、活用し易くなる。しかし、医療情報においては、氏名や患者ID等の個人特定情報を除く無名化では、データセットの構成に依存するが、一意性が残り、データの対象集団が特定されれば、再特定化が可能になり得る。特に、一意性のあるデータが特異であると、容易に再特定化ができる。
- 高い匿名性のためには、一意性を除去することが望ましいが、そのためには元データを加工することが必要となる。その結果として、元のデータ特性が大きく変わってしまえば、匿名加工のデータを活用する意味がなくなる。匿名性の高さと元のデータ特性からのずれの大きさはトレードオフの関係にある。
- 2017年5月の個人情報保護法の改正では、匿名加工情報という考え方が明示され、「個人情報の保護に関する法律施行規則」において、匿名加工情報の作成の方法に関する基準が示された。しかし、その基準は一般的のものであり、対象とするデータと利用目的ごとにどの程度の加工が必要かは検討しなければならない。

【研究目的】

i. 患者が増加している生活習慣病を追究するためには、病気が発症する前の健診データが重要である。そこで、健診実施施設が健診データ利活用についてどのように認識しており、健診データの外部提供においてどのような点に不安を持っているか、また匿名化技術についてどのような認識にあるかを明らかにする。

ii. 複数の医療機関からデータ収集してデータ分析することを想定して、3大学病院から健常人に近いデータを集め、匿名性の高さと統計量のずれの関係を調べ、適切な特異性除去・一意性除去が可能かを検討する。

ポスター2


【研究目標】

1-1 健診実施施設が保有する健診データをどのように二次利用しているかを調査する。

1-2 健診実施施設が保有する健診データを外部機関へどの程度提供し、提供する際にどのようなことを条件としているかを調査する。特に、匿名化についてどのように認識しているかを把握する。

2-1 データ変数ごとの匿名化による基本統計量（平均値、分散、四部位値）と頻度分布への影響を、特異性除去と一意性除去の方法を変えて評価する。特異性除去では、(1)平均値±3SD、(2)2.5percentileから97.5percentileに収まらないデータを外れ値とする2つの方法で評価する。一意性除去では、k-匿名性を確率的な指標に拡張させたPk-匿名性を用いる方法で、k値を複数設定し、特異性除去方法および一意性除去方法による影響の違いを評価する。

2-2 データ変数ごとの匿名化では相関性の強い変数の組で見ると一意性除去が出来ていないことが生じる。そこで、相関性の強い変数の組で特異性除去を平均±3SDで行い、一意性を単純な乱数ノイズ付加の方法で行い、それらの影響を評価する。

【研究方法1】

健診実施施設に対する医療情報利活用の状況調査と匿名化に対する意識調査は以下の手順実施した。

① 調査対象施設：日本医療・健康情報研究所がホームページ上で公表している533の全国健康診査実施施設から、都道府県別の施設数に応じて300施設を抽出し、調査対象施設とした（300施設に限ったのは調査費の制約による）。

② 調査回答者：調査の目的を説明した文書を添付し、施設情報管理者に回答を依頼した。したがって、実際の回答者は事務職が76%であったが、施設情報管理者の指揮下で回答されたものと認識した。

③ 調査実施期間：平成28年7月18日から8月31日

④ 倫理的配慮：回答票には施設名や回答者名の記載を求めず、無記名とした。さらに、調査依頼文には、回答は施設の自由意思によること、調査協力の了承は調査への回答をもって得たものとすることおよび結果公表に当たっては対象施設名を公表しないことを明記した。

今回、共同研究者の愛媛大学の木村先生が、その匿名化に対する技術を持っておられましたので、それを使ってやってみようということになりました。

研究目的ですが、健診のデータは医療データと違って、全ての人が受けるデータとしてあるわけなので、非常に膨大なデータになります。病気になった状況だけでなく、病気の前の情報も大事ですので、その経過を見ていくという意味では、こういったデータも使えないとならない。そこで今回、病院のデータそのものというよりは、健診データを想定しながら分析してみようということでしたので、健診データを持っている健診実施施設が、こういう匿名化、あるいはデータの二次利用に関してどういう意識であるかということを、一つ、確認しました。

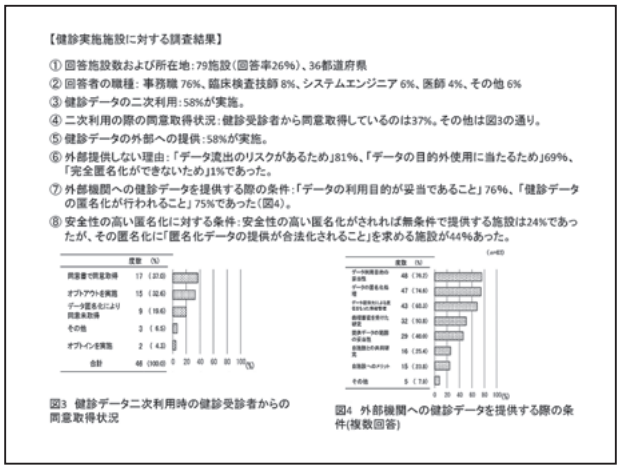
その上で、データの匿名化の評価に関して、この研究に参加した3つの大学からデータを取ってきて、それを分析しました。一つは特異な部分のデータを排除することと、一意性を排除することです。一意性をどこまで排除するかということは実は難しいのですが、医療データの場合はデータ項目がたくさんあるので、原理的に言えば、組み合わせをやっていくとほぼ無限な形の組み合わせになって、すぐ一意性は出てきてしまうのですが、その出方によって除き方が若干違うのかなと考えています。

それで、その最初の部分の健診実施施設での調査ですが、今、健診施設が全国で533施設あります。その中から300をランダムに選びました。

【ポスター3】

結果として、回答を得られたのは少なくて、79施設です。26％なので、この回答は若干バイアスがあるかもしれません。主に、健診の二次利用に関して関心のある施設が返してきている可能性はあります。

ポスター 3



その中で、健診施設自身でも二次利用をしているところは60％ぐらいあって、さらに外部に提供しているか…大学とか研究機関とかに提供しているかいうと、これも同じ60％ぐらいの施設が提供していました。ただ、逆に、していないところももちろんあるわけですけれども、そこでは一つの大きな理由としては、データ流出に関する心配があるのと、それから、そもそも同意を得ていないということなので、そういう場合だと、匿名化したとしても目的外使用になるのではないかと懸念されているところがとても大きなウエートを占めていました。

それから、匿名化をしたらどうかについて。先ほど言いましたように、単に無名化ではなくて、個人が再特定できないような匿名化をした場合どうかということに関しては、24％ぐらいの施設は、その時点でOKという回答でしたが、プラス、法律的に合法化されていればいいであろうというのが44％です。これは残りの44％なので、実質的には半分ぐらいです。そういう状況になっています。

その意味では、今回、個人情報保護法が改定されて、匿名化されたデータを使うルールが明確化になってきたということは、非常にいい流れなのかなと思います。そういう形の運用ができてくれば、今後もう少し利用が進んでいく可能性はあるのかなと考えています。

【ポスター4, 5】

あと、実際に再特定化されないようにするためには、データを若干いじらないとそういう状況はつくれません。まず一つは、特異性です。全体データから見たときに外れているデータはやはり特異ですので、それを外す。外し方としては非常にシンプルなのですけども、集団の平均、プラス・マイナス標準偏差の3倍の範囲で、それを越えるもの。あるいはもうちょっと緩いやり方としては、下から2.5％より小さいものは特異、それから97.5％を超えるものは特異という扱いで、その二つのやり方をしたときにどうなるか。

それから、それは単に特異なものを外しただけで、一意性をまだ排除していないので、一意性を排除するためにランダムにデータにノイズを入れるというやり方をしました。入れ方がまたいろいろとあるのですが、今回は、一つは、ラプラス関数という確率関数で確率的なノイズを入れるというやり方です。そうすると、同定される個体数(k)の期待値が求まりますので、それを変数として見ていき、大きくしていったらどうなるかということを評価しました。

ポスター4

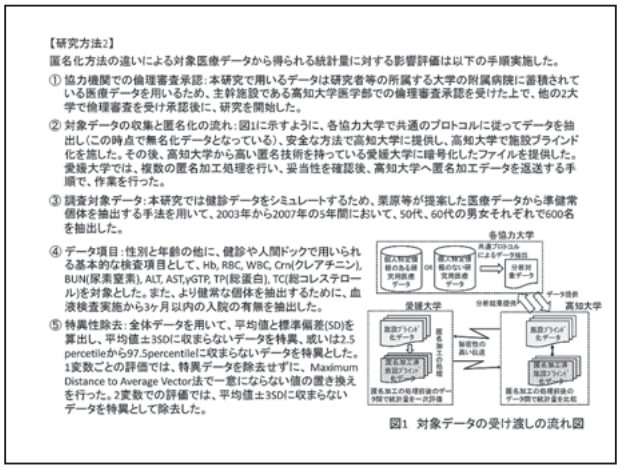


図1 対象データの受け渡しの流れ図

ポスター5

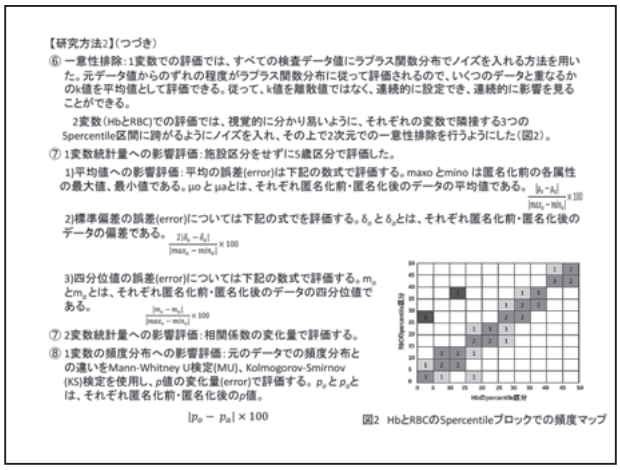


図2 HbとRBCのSpercentileブロックでの頻度マップ

それからあと、相関性です。医療データの場合、結構2つの変数が強い相関を持っている場合があります。そうすると、それぞれはばらけたけれども、相関性から外れてしまってずれているというものは、ある意味で、本当は特異値を外したはずなのだけれども、特異性がまだ残っているということなので、それを今回のデータの中では、特に、赤血球数とヘモグロビンの値で評価してみました。

【ポスター6, 7】

今回した統計の方の結果は、1変数で見ていったときに、平均値±3SDで取る場合と2.5％から97.5％の間だけ使うというやり方には、あまり差はないということが分かりました。

赤血球数は割と分布としては正規分布に近い分布ですが、それに対してγGTPという検査は正規分布から外れていて、テールをざーと引くような分布です。そうすると、そういうところだと、結構、外れというか、長くデータがあるので、その辺にノイズを入れると、例えば、平均値でも結構ずれてしまうということが分かりました。それから標準偏差もかなりずれてしまう。

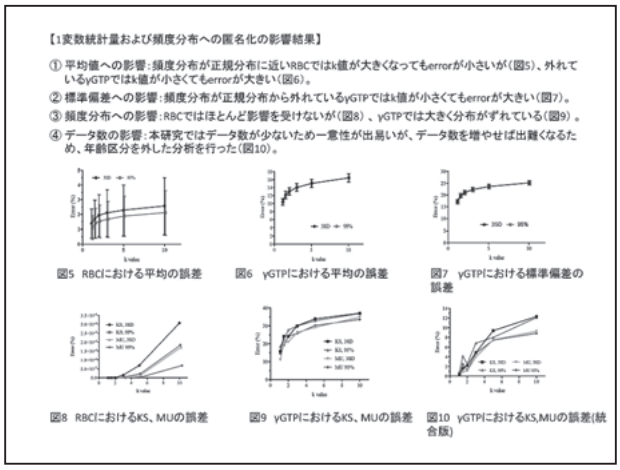
あと、見たものは分布としてどうかということで、コルモゴロフスミルノフの検定とマンホイットニーのU検定をやったのですが、それで見たときでも正規分布から外れた分布はなかなか扱いが難しいなというのが分かりました。

あと、相関性に関しては、今回、簡単な部分で見たのですけれども、5パーセンタイル単位に切った上で、その中でランダムに混ぜるということをやりました。それでも相関関係から大きく外れるのが出てきて、そこの値を削除することをやったときには、若干、相関性がずれてくる。3％弱ぐらい相関係数にずれが出てくるという結果が得られました。

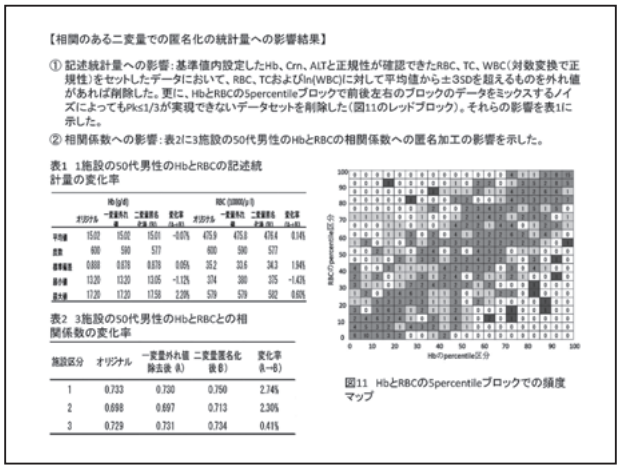
【ポスター8】

以上の結論として、一つは健診のデータを利用していく上では、今後は個人情報保護法に則った形で匿名化をするということと、それを利用するときに利用施設に対してちゃんと説明することをやっていけば、進

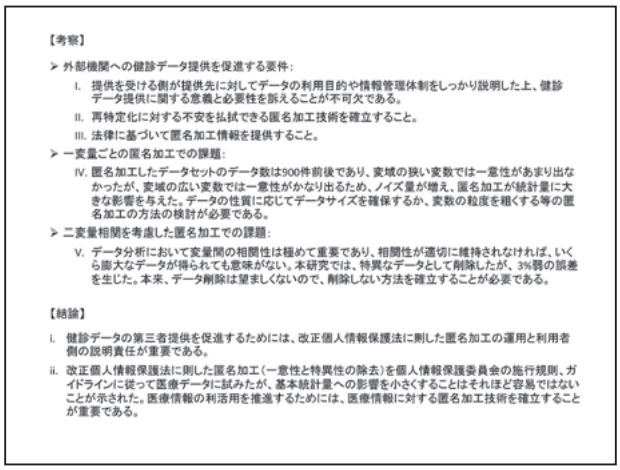
ポスター 6



ポスター 7



ポスター 8



んでいくのではないか。

それから、匿名化の加工をするというところは意外と簡単ではなくて、もうちょっといろいろと工夫をしていかないと、統計量に影響が出てしまうというところが今回の結論です。

10部ほど報告書を用意してきましたので、関心ある方はどうぞお持ち帰りください。

質疑応答

座長： ちょうど個人情報保護法が改正されたということで、タイムリーでもありますし、また、ビッグデータの活用を医療領域でも何とかしようという動きもありますから、そういう意味でも非常に貴重なご報告だったと思います。
ところで、法改正がなされたばかりですので、ある意味では混乱もあると思うのです。先生方は医療現場で、どのようにうまく新しい法律に整合するように活用する工夫をされているのか、その辺をご説明ください。

栗原： 実は、改正されたのですが、大学病院などアカデミックのところは適用外であります。しかし、一応、準じる形でやっています。今回の発表は同意を得ない・未取得でという言い方をしましたが、例えば倫理委員会などでの扱いに関しては、基本的には同意を取る方向が強くなってきているのかなと思います。あと、匿名化して使う場合にはどういうデータが対象になるかということを明示、公開して、それが嫌だという方はブロックすることができるようにしなければいけません。匿名化して研究するときには、運用としては、ちゃんと事前に大学の附属病院のホームページで、どのデータを使います、どういう患者さんが対象になります、ということをやりつつあると思うのですが、これが今、全国的にどこまで普及できているかはまだ把握はできていません。

座長： これは全国の課題でしょうね。いろいろな形で法律のほうでも今、議論をしているところですし、それ以外の臨床現場でもこういった問題にどう対処すべきかを、いろいろ検討していると聞いております。先生のご研究がより具体化していくことを期待しております。