

紹介

医療・健診データの科学分析の重要性と個人情報に依拠した公的データの研究の難しさ

土岐 博

(大阪大学キャンパスライフ健康支援・相談センター特任教授)

目次

- I はじめに
- II 認知症になる原因を知りたい
- III 大阪府 KDB データは宝の宝庫だが全貌を掴むのは大変
- IV 個人情報であるゆえの匿名化とデータ間の突合の難しさ
- V アスマイルのための AI を作る
- VI 医療健診データの分析の難しさ
- VII ハッシュ化された公的データのあるべき姿
- VIII 大阪大学での KDB データ
- IX おわりに

I はじめに

大阪大学に 2018 年度末に導入された大阪府の国保データベース (KDB) の導入経緯と、ビッグデータである医療・健診データを分析する際の経験を記す。我々のグループは純粋に科学研究を行うことを目的として KDB データを導入した。その分析には予期せぬことが多くあったのでその経験も書いておきたい。3 年目からは大阪府が運営する健康アプリであるアスマイルに生活習慣病発症確率を予測する人工知能 (AI) を作ることを依頼されたこともあり、KDB データの利用は実際の社会にも役立つ成果を作り出すことができた。最後に個人情報に依拠した公的データを扱う際の

共同研究の難しさにも触れたい。

II 認知症になる原因を知りたい

私は 2010 年の 3 月に大阪大学を定年退職した。現役時代には原子核物理の理論研究に埋没していた。退職後はすべての時間を自分のために使えるので、退職する際に半分の時間は好きな原子核研究に使い、半分は社会貢献に使おうと考えた。

その頃に社会問題になっていた認知症について社会貢献の時間を使うことにした。退職して数年後に阪大内に全学の認知症横断プロジェクトを立ち上げた。3 年間に月 1 回のペースで大学内の専門の認知症研究者のセミナーを企画した。参加者も自由に質問・コメントできる場を設定して勉強した。毎回大学の内外から 50 人くらいが参加する活気あふれる談話会だった。良い話がいっぱいあったので、認知症横断プロジェクトの主催者が編集して『ほんとうのトコロ、認知症ってなに?』という本を大阪大学出版会から出版した。

認知症は高齢になり認知機能が衰えていくことにより誰にでも起こる老人病であるとの結論は得たが、症状の重い人や軽い人がいる。その差はどこから生まれるのかに興味を持った。日常生活習慣を反映している血糖値などの数値や問診結果が健診データには記録されており、その差を生み出す原因が隠されているのではないかと考えた。

そこで、そのセミナーに参加していた市役所の職員の人に健診データを分析したいのでデータが欲しいとお願いした。すぐに大阪府国民健康保険団体連合会に連れて行ってもらった。連合会の職員の人たちに見事に整備されている KDB データの内容を紹介していただいた。大阪府全体の国民健康保険被保険者（国保）の健診及び医療データが 2012 年度から蓄積されていた。しかも、大阪府のはほぼ全市町村のデータが揃っているということであった。データ使用料は必要ではあるが、大阪大学に提供することができるということであった。そのときにデータ使用料は個人が特定できないように加工するために 500 万円くらい必要であると言われた。退職していて、しかも分野の全く違う私には支払う力はなかった。そこで、連合会と大阪府にお願いして、国にデータ分析のプロジェクトの申請をしていただくことにした。データを調査分析する事業としてこの費用を国に申請した。

2018 年の年末に申請が受理されたという知らせが来た。本当に短期間ではあったが、その時点から大学内のいろんな障壁を乗り越えて、無事データを導入できたのは 2018 年度の最終日であった。その間に連合会では KDB データの匿名化を行っていた。それでも最終日にデータを受け取りに行くまで連合会の人たちは心配で仕方がなかったということだった。最後の一押しは国が認めているプロジェクトだからということであった。

Ⅲ 大阪府 KDB データは宝の宝庫だが全貌を掴むのは大変

私自身、このデータを分析するということはどういうことなのかを全く理解していなかった。データ量は全部で約 1Tb（テラバイト）あった。しかもデータを使うということでもつけられたプロジェクト名は「生活習慣病予防・健康寿命延伸を目指した調査分析事業」であった。さっそく仲間を集めた。医学の先生、看護の先生、それと物理の先生である。医学の専門知識は絶対に必要だが、大量のデータを扱う技術も必要だった。物理は私の専門だが、このデータに興味を持ってくれた 2 人の医学物理の先生にも入ってもらった。総

勢で 9 人の先生方とその学生さんたちであった。予算ゼロのプロジェクトなので、すべての分析や、議論などは参加者の研究時間と研究費で行った。その意味ではこのデータ分析の動機は科学的に有用な知見を得て、科学論文として発表することであった。

医療摘要データはすべての薬や入院などの個人の医療の記録が記載されている膨大なものであった。薬などは番号で与えられており、その番号が何を表すのかをすべて調べる必要があった。一方で健診データは身長や体重や血圧などの項目名で数字が入っていることでとっつきが易しいことで、まずは健診データの分析に時間を使うことにした。ところが健診データには生年月日と性別が入っていなかった。それらがあるのは台帳と呼ばれるファイルであった。そこで台帳の分析を始めると、ハッシュ化された被保険者がいつ国保に入り、いつ国保から出ていったのかを正確に把握するのが非常に難しいことがわかった。さらにはすべての人が台帳に登録しているものだと思って突合してみると、かなりの人のデータが存在しないことがわかった。台帳は市町村の転出や転入があるとき、介護保険の変化があったときには更新されるという仕組みであり、国保や後期高齢者保険制度（後期）に 2012 年以前に加入しているが、転入転出などの変化がない限りはその人はハッシュ化された台帳には入っていないことがわかった。それが 11 月ごろで、連合会にその情報の提供をお願いして、最終的には欲しいデータが揃ったのは 2018 年度の終わりの 2 月ごろであった。1 年がこんなことで過ぎていった。無駄ではないが、かなりの努力が必要な作業であることは間違いない。

最初に KDB を導入したときには台帳は非常に大きなファイルであった。解析を進めていくと、台帳は被保険者に何らかの変化があるときに更新されることがわかった。介護保険の要介護度が変化するときも更新されている。最初の段階で理解できなかったのはこの更新は過去にまで遡って行われるということであった。このことが台帳を膨大なものに行っている理由だったので、最終的な台帳は過去の記録にまで影響を与えることを避ける

ようにしてもらった。予見を持ってデータを分析するのだが、そのことで台帳が理解できない期間が非常に長いものになった。

IV 個人情報であるゆえの匿名化とデータ間の突合の難しさ

KDB データを阪大に移管することの不安が最後までつきまとったことは先ほど述べたが、移管された KDB データでは個人を同定できないようにするためにいろんなところが落とされている。生年月日は生年だけが残されており、住所は当然落とされている。さらには国保に入った期日は入っているが、国保から抜けたときの情報が入っていないこともあり、ある被保険者がある市にどれだけの期間住んでいたかを正確に把握することが難しい。ある被保険者がある市から別の市に移った場合にはハッシュ化番号が変更するようになっていた。転居の多い人を追跡するのは難しく解析を困難にしていた。ある期間にどれだけの人が被保険者になっているかを同定するのは難しい。さらには多くの国保の被保険者は 75 歳の誕生日に後期に移行する必要があるが、国保でどのハッシュ番号の人が後期でどのハッシュ番号の人なのかを対応させることは難しいこともわかった。このことで、75 歳である人の健康や医療履歴を切り離して扱わざるを得なかった。男性の健康寿命は 72 歳で平均寿命は 82 歳、女性の健康寿命は 75 歳で平均寿命は 87 歳なので、健診データに現れる老化による体の変化と医療や介護の変遷を追跡することが困難なことは非常に残念なことである。

KDB データが導入されてから 7 年以上が経過したが、ある市町村で何人の人がある日にいるのかを確定することは非常に難しいことがわかった。これは母数を確保するのに難しさがあるということである。このことは科学論文を発表する立場では非常に辛いところである。その意味では KDB データの所有者である市町村との直接の共同研究の柱がたてば、この点は解消される可能性がある。市町村でデータの中身を正確に掴むような研究に興味を示す市町村は少ない。KDB デー

タではデータの保有者は市町村であることと、ビッグデータでは匿名化は避けては通れないところが研究の難しさの原因になっていると考える。

それでも 2 年目でなんとか健診データを使って、それぞれの健診項目の間の因果関係を導出する研究を行うことができた。これは割と早い時期に論文として発表することができた。この論文は我々のグループをまとめるのに大いに役立ったと考えている。

V アスマイルのための AI を作る

KDB データが導入されて 3 年目になるが、大阪府から健康アプリのアスマイルに生活習慣病を予測する AI を開発するプロジェクトを委託していただいた。それまでの 2 年間で健診データはよくわかっていたし、それぞれの薬名がどんな病気に使われるものなのかの分析もできていた。我々のプロジェクトとしても AI を作ることは視野に入っていた。その意味では 2 年目の終わり頃にアスマイル AI を作るプロジェクトを得たことは嬉しい限りであった。

それまでは、筆者の学問的興味から手弁当での研究だったが、このプロジェクトのおかげでシニアスタッフ 1 名とジュニアスタッフ 2 名、さらには大学院学生をデータ分析で雇うことが可能になった。研究室も豊中キャンパスの保健センター内に持たせてもらうことができて、このプロジェクトに専念できることになった。

我々のグループには AI を作る研究に携わっている研究者はいた。しかし、学外の先生であり、招聘教授になっていただいて助言をいただくことで対応せざるを得なかった。その先生方は直接にデータを扱うことができず、プロジェクトに参加している阪大内のスタッフと学生で独自で AI の勉強を行い、実際のデータを動かしてみ、どんな答えを得ることができるかを試行錯誤した。1 年の期間はあっという間に経過した。最初はニューラルネットワークを使うつもりで勉強していたが、たくさんのデータを扱う際の計算時間が長すぎることもわかり、1 年を経た頃には

lightGBM という決定木 (decision tree) の方法を採用することにした。この方法はある条件に対して yes/no で判断しながら領域分けしていくので計算機にとっては得意な方法であり、非常に短時間で答えを得ることができた。AI の内容としてはしばらくは、病気と断定するのにそれぞれの病気の治療に特有の薬の服用を使っていたが、それぞれの生活習慣病の学会が発表する病気発症数の数分の 1 の患者数にしかないことも判明した。本当に試行錯誤が続いた。苦しい 1 年目であったと記憶に残っている。1 年目の報告書は最後まで何を書くかが決まらず、最後の 1 週間でなんとか書き上げた。

2 年目に入った頃から方針が見えてきた。ジュニアスタッフも新しいメンバーを投入することができて、AI 生成に専念することが可能になった。病気が発症したかどうかを決めるのにそれぞれの学会が発表している病気発症の診断基準を使うことにした。糖尿病なら HbA1c が 6.5% 以上とか、高血圧なら収縮期血圧値が 140mmHg 以上とするなど、学会が決めている病気発症基準を使うことにした。そうするとすべての数字を健診データから得られることもできるので扱うデータ量もかなり抑えることができるようになった。スタッフの数が増えたことと、研究に使えるワークステーションの数が増えたことで、かなりスムーズに研究を進めることができるようになった。ほぼ 6 月くらいの段階で、糖尿病、高血圧症、脂質異常症の 3 大生活習慣病の 3 年以内の発症確率を計算する方法を決定することができた。最終盤ではないが非常に良い AI ができる自信を持つことができた。最終盤のためにはできるだけ多くのデータでチェックできるように最後まで AI 内部の改良を行った。

さらには、個人情報がいかりと守られる必要がある、アスマイルに登録されている健診データを使って、今回製作した AI で結果を得ることができるようになる必要があった。そのために AI をクラウド上のコンピュータに搭載した。アスマイルユーザーの健診データが入力されて、結果を得るようにし、データは蓄積されないような仕様にした。最終結果は個人のアスマイルにアウト

プットが表示されるような仕組みを作り上げることにした。この辺りも全くの素人だが、なんとか保健センターに出入りの業者に委託をすることで実現することができた。やっとアスマイルに AI を公開したのは 2 年目の 12 月 10 日のことであった。実際にアスマイルに自分のデータを入れてそのアウトプットを受けることができたのは本当に感激の一瞬であった。大変だったが、すごいことを成し遂げたという思いを持つことができた。

VI 医療健診データの分析の難しさ

KDB データに携わる研究者はいろんな分野の人たちである。個人を特定できないくらいに個人情報情報を排除したデータと言えども元は個人情報であるデータの取り扱いには慎重を期する必要がある。ハッシュ化された KDB の元データは 1 つのワークステーションに入れておき、ハッシュ番号をあえて、別の番号に加工したものを作って、研究場所を特定し、そのデータの入っているワークステーションの数をできるだけ少なくして、その場所に物理的にいる人だけがそのデータを分析できるような仕組みを構築した。これは研究に大きな制限を加えるもので、空き時間で研究するというようなことができなかった。これは研究のスピードを大きく遅らせる要因となった。

健診データは項目もわかりやすく、性別と生年月日が入れば分析しやすいデータであった。KDB データには医療摘要というファイルがある。ここには被保険者の医療行動のすべてが反映されている。どんな薬が処方されたかも記録されるし、入院しているかどうかもわかる。さらには医療費がどれだけかかっているかも記録されている。細部に渡って書かれているので、医療行為を追跡しながら、病気の進行や回復を分析したいと思ってはいるが、多大な時間を要する研究でなかなか論文までには到達するのは難しい。データ分析を難しくしている要因の 1 つに傷病名が使えない (使いにくい) ことがある。医師によれば傷病名と使う薬が関係しないと使えないということもあり、その病気ではない場合にもある傷病名を記述するということがあるので、傷病名は使えないとい

うことである。

分析しやすいのは医療費である。もともと、それぞれの保険者が医療費の一部を負担する必要があるので、この数字は非常に正確である。実際に1人当たりの医療費や、さらには年次変化を分析すると厚労省が発表している年間の総医療費と一致することも確かめることができた。それとしっかりしているのは要介護度の変遷である。これらの情報は割と早い時期に把握することができた。2020年の段階では総医療費は健康保険に登録されているだけで約50兆円になる。男女別年齢別で1人当たりの医療費も計算できるし、これからの人口分布もわかっているのだから、30年後の医療費・介護費を計算することができる。超高齢化の影響もあるが、医療技術の発展から医療費そのものも高騰しており、30年後には約100兆円に達することが計算される。これは政策だけで解決できる問題ではなくて、本当にKDBデータを分析して、何が病気の原因になっているか、どうすれば病気を予防できるかをしっかりと把握する必要があることを示唆している。

Ⅶ ハッシュ化された公的データのあるべき姿

KDBデータは素晴らしいビッグデータであると言える。個人情報が満載されているデータであることで、どうしても多くの情報を落とす必要がある。おそらくはすべての情報が入っていれば、割と苦労なしに必要な情報を得ることができたと考えている。ところが多くの情報が落ちてしまっていることと、ハッシュ化の方法がうまく個人を追いかけるようになっていないこともあり、誰がいつどこにいるのかを同定することが難しくなっている。その意味では個人情報を落とす際のアルゴリズムとハッシュ化の方法に実際に分析する科学者が入り込むことが必要だと思う。つまりはハッシュ化したデータで分析し、問題があればハッシュ化の方法を改善して、ハッシュ化データを作るということを繰り返す必要がある。この辺りはKDBデータを管理する国が仕組みを作り上げる必要があると考える。

Ⅷ 大阪大学でのKDBデータ

アスマイルのAI製作の委託期間も終了し、今は若い研究者もそれぞれが新しい勤務地に移動して、今は大阪大学の教員とその教員を取り巻く大学院生やポストドクターが研究者としてデータ分析を行っている。ゆっくりだが、興味本位での研究は続いている。

KDBデータも2012年度のものから2022年度のものまで10年分のデータが保健センターに保管されている。研究プロジェクトの期間としては2025年度いっぱいには約束されている。おそらくは研究プロジェクトの更新が行われるものと思っており、多くの新しい研究者が参入してくださることを願っている。

Ⅸ おわりに

私は物理の理論研究者である。物理のデータはすべてが公開されている。この事実のありがたさを今回のプロジェクトで身にしみるように理解できた。物理研究では自然探求の興味が無限に研究者の時間を使わせてくれる。だからこそ、世界の共同研究者と時間を気にせずに研究連絡をしながら研究を進めることができる。そのことのありがたさがわかった。

ところが、個人が同定できないようにまで個人情報が落とされており、しかもハッシュ化という個人番号のみが全体のデータを突合できるようになっているという意味では、個人情報満載のデータを興味本位で研究するというのは本当に難しい。そのことで、物理では世界の研究者と共同研究できるが、個人情報からみのデータを掘り所とした興味が動機になっている研究は本当に難しいと言える。その意味では現在のように大学内で興味を持っている研究者が、データのある部屋に入りして研究を行うということで満足する必要があると思う。

私はいつも「データは愛情を持って見ないと何も語ってはくれない」とよく発言している。すべてのデータは自然の情報を運んでいる。自然を知

るにはデータを読み解く必要がある。数学やコンピュータの技術は修得している必要があるし、時間も無制限に使う必要がある。そのような科学者の努力があってこそデータは語り始めてくれる。

医療・健診データの科学分析のあるべき姿は、国や地方自治体がスポンサーになり、ハッシュ化データを作るところにも科学者が携わることができるようにした上で、ハッシュ化データのある程度の制限はつけるもののでできるだけ公開し、多くの科学者がデータ分析を行えるようにすることだと思う。このようにすることで初めて「生活習慣病予防・健康寿命延伸を目指した調査分析事業」のような表題のプロジェクトは本当に欲しい答えを得ることができると考えている。

最後に今の段階での私の希望を述べたい。アスマイルに搭載した三大生活習慣病の3年以内の発症確率を予測するAIは非常に優れたソフトである。内部検証、外部検証、経年検証のすべてで抜

群の性能を出している。現在は大阪府の国保会員しか使うことができない。大阪府と厚労省が協力して、このAIを日本全体の人たちが使えるようになることを心から望んでいる。そのことにより、国民のすべてが生活習慣病の発症確率を自らで知ることができて、自律的に生活習慣を改善する努力をできるようにしたい。自律的に生活習慣病の発症を抑えることや、遅らせることができると考えている。生活習慣病はすべての重病の入り口にあるので、自らの生活も健康的に送ることができるし、医療費という意味でも抑制することができると考えている。

とき・ひろし 大阪大学キャンパスライフ健康支援・相談センター特任教授、大阪大学核物理研究センター名誉教授。主な論文に“Causal Relations of Health Indices Inferred Statistically Using the DirectLiNGAM Algorithm from Big Data of Osaka Prefecture Health Checkups,” *PLoS ONE*, Vol. 15, No. 12, e0243229 (共著, 2020年)。原子核物理理論専攻。