エビデンスに基づく政策形成と統計学

上越教育大学大学院学校教育研究科・学校教育学系
准教授　奥村太一

1. はじめに

近年，政策形成に科学的な視点を盛り込むことに注目が集まっている。従来，政策といえば政治家や官僚，有識者などの意見のすり合わせによって作り込まれるものであった。すなわち，政策の方向性や具体的な内容は，人間の様々な利害や経験則，信念や思惑によって左右される。このやり方の根本的な問題は，その政策によって人々の行動がどう変容し，その効果がどのような形で表れるのかを示す客観的な根拠に欠けていることである。どのような将来が人々に確約されるのか不明確なままでは，後々その有効性を検証することはもちろん，仮に失敗だったという結論が出たとしても，その原因を特定することすら難しい。意見のすり合わせによって作られた政策をいきなりフルスロットルで実行するのは，言ってみれば壮大な賭けに打って出るようなものである。

政策の実施には莫大な費用がかかる。吉と出るか凶と出るか明らかでないものを一気呵成に推し進めるよりも，まずは限られた範囲内で試行的に実施し，効果が明らかになったものを本格的に実施するという段階を踏むほうが，明らかに経済的に理にかなっている。このように，主観的な「意見」よりも客観的な「事実」をよりどころとして政策作りを進めることを「エビデンスに基づく政策形成」(evidence-based policy making) という。ここでいう「エビデンス」とは科学的根拠のことである。特に，ある変化が確かに特定の働きかけによって生じたという原因と結果のつながり（因果関係）の裏づけとなるものを指すことが多い。諸外国では，税金の無駄遣いを防ぐために多くの政策分野でエビデンスを得るためのフィールド実験[1]が行われている。わが国ではそもそも「エビデンス」という概念自体があまり浸透しておらず，産学官の先進的な取り組みとしてごく少数の試みが行われているにとどまっている[2]。

2. エビデンスに基づく医療

公的な意思決定においてエビデンスを求める動きは，医療分野が先んじていた。1992 年にカナダの医学研究者ガイアットらは「エビデンスに基づく医療」(evidence-based medicine: EBM) という提言を行った。これは，医師の臨床経験ばかりに頼るのではなく，検査データを過去の臨床研究から得られた科学的知見と突き合わせて最も有効性の期待できる治療（標準治療）を選択するべきだとする考え方である。患者にしてみれば，医師によって治療方針が一貫しないという事態を避けることができるし，医師にしてみれば，専門分野の細分化が進み誰もがすべての疾患について専門家とはなりえない時代において，客観的なガイドラインがあったほうが診療を行いやすい。また，国にしてみれば，少しでも有効性の高い治療を第一選択としてもらった方が医療費を抑えられ財政面で助かる。このような事情もあって，EBM は（実態はどうあれ）現代医療における基本的な行動指針となっている。

臨床研究の結果を治療の拠り所として用いるのならば，それらがあちこちの医学雑誌や報告書に散逸しているのでは使い勝手が悪い。そこで，様々な臨床研究の結果を一箇所に集約し，研究成果を統合してエビデンスを強化する作業が行われるようになった。イギリスに本部を置くコクラン共同計画（Cochrane Collaboration）は，このために設立された国際的な非営利団体である[3]。ここでは，世界各地で行われた臨床研究の結果を

[1] 人々の生活の場を利用した実験のこと。実験室のように管理された環境で行われる実験は実験室実験と呼ばれる。

[2] 朝日新聞 2017 年 10 月 26 日「その政策，科学的に有効？根拠に基づく立案 日本でも広がるか」などを参照。

[3] http://www.cochrane.org/ja/evidence

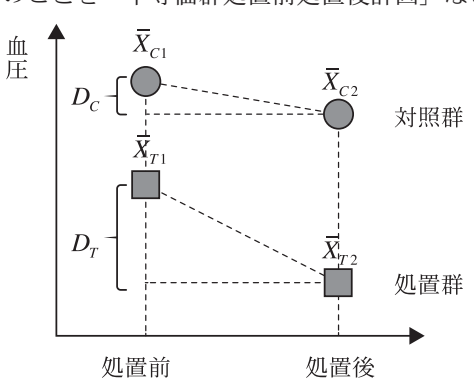
収集し，それらの質を評価して選別された結果を統合し，治療方針を決定するための様々なガイドラインを提供している。

ここで重要なのは，とにかくデータを集めて分析さえすればどのような臨床研究の結果も等しくエビデンスとして扱われるわけではないということである。コクラン共同計画では，最も強いエビデンスを与える臨床研究は無作為化比較試験（randomized controlled trial: RCT）に基づくものであると規定している。RCT とは，治療を受ける群（処置群）と受けない群（対照群）に患者を無作為（ランダム）に割り当てて，群間で結果の違いを比較しようというものである[4]。無作為割り当てが行われていれば，2 群に割り当てられた患者は治療の有無以外はすべて等価であることが期待されるため，処置群と対照群の結果の差をそのまま治療の効果とみなすことができる。これについては，統計的検定のロジックとあわせて前回のコラムでも説明した[5]。

3. エビデンスとしての質

RCT の強みは，その名が示す通り患者を複数の群に無作為に割り当てることにある。これによって，例えばたまたま症状の軽い患者が処置群に多く割り当てられる，といった偏りが生じる可能性を確率的に評価できるためである。しかし，患者が希望する治療を受けられることが確約できないという倫理的問題や，多くのスタッフと協力し綿密な管理のもとで実施する必要があるという手間を考えると，RCT を行うことが現実的でない場合もある。そのため，治療の有無によって結果を比較はするが，患者の無作為割り当ては行わないという妥協案が採用されることもある。

例えば，治療を希望した患者は処置群へ，希望しなかった患者は対照群へ割り当てれば，先述の倫理的問題は生じない。あるいは，治療を行う施設が整っている病院の患者を処置群，そうでない病院の患者を対照群とすれば，管理の手間もかなり省くことができるだろう。しかし，この方法では処置群と対照群が等価であることは期待できない。なぜならば，前者の例ではより症状の重い患者が治療を受けることを希望するかもしれないし，後者の例では先進的な治療が行える病院にかかる患者の方が健康に関して意識が高く全体的に症状が軽いといった違いがあるかもしれないからである。そのため，処置群と対照群の治療期間後の結果を単純に比較しただけでは，その差が治療によって生じたものなのか前から存在していたものなのか区別できない。そこで，治療開始前の段階で 2 群にどれくらいの差異があったのかを考慮して結果を比較するということが行われる。下の図を見てほしい。今，高血圧の治療効果を検証したいとする。処置群と対照群の血圧について，処置前の平均を $\bar{X}_{T1}$，$\bar{X}_{C1}$，処置後の平均を $\bar{X}_{T2}$，$\bar{X}_{C2}$ とする。このとき，処置群の変化を $D_T = \bar{X}_{T1} - \bar{X}_{T2}$，対照群の変化を $D_C = \bar{X}_{C1} - \bar{X}_{C2}$ とすると，この差 $D = D_T - D_C$ を治療の効果と見積もるのである。これは $D = \left(\bar{X}_{C2} - \bar{X}_{T2}\right) - \left(\bar{X}_{C1} - \bar{X}_{T1}\right)$ であるから，処置後に見られた群間差から処置前に存在していた群間差を差し引いていることになる。これを，「差の差分析」といい，この例のように不等価な群の処置前後の変化を比較する研究計画のことを「不等価群処置前処置後計画」など



[4] 実際には，「この治療はきっと有効だ」という期待が結果に影響することを避けるため，患者本人はもちろん，実際治療にあたる医師や看護師にもどの患者がどちらの条件に割り当てられたかということは知らされないことが多い。これを二重盲検法という。薬を服用する場合には，処方の有無によって割り当てが特定されないよう，有効成分は全く含まれていない偽薬（プラセボ）が用いられる。もし新薬に副作用が予想される場合は，副作用の出現の有無によってどちらの条件に割り当てられたか感づかれる可能性があるため，副作用に似た症状を引き起こす成分のみを含む偽薬をわざわざ用いることもある。

[5] http://www.jikkyo.co.jp/contents/download/9992658085

という。

先ほど，これをRCTに対する「妥協案」と書いた。それは，差の差分析で処置前の違いを考慮したとしても，得られた結果はRCTに比べてエビデンスの点で劣るからである。差の差分析によって治療効果が正しく評価できるためには，もし処置群が治療を受けなかったとしたら，処置前から処置後への変化は対照群と同じであることが保証されていないといけない。例えば，先進的な治療を行える病院はそうでない病院に比べ医療技術や看護の質も全般的に高いかもしれない。すると，処置群は対照群より症状が軽かっただけなく，恵まれた環境に置かれているため治療を受けなかったとしても対照群より良好な経過をたどっていたかもしれない。こうした可能性を排除できない限り，治療を受けることによって生じた変化が純粋にどれくらいあったのかを見積もることはできない。

このように，無作為化という要件をひとつ外すだけで，治療の有無以外に結果の違いを生み出したかもしれない様々な可能性を考慮に入れる必要が出てくる。そのため，複数時点で観測を行ったり，結果に影響を及ぼしそうな他の要因についても情報を収集したりと，RCTに比べてデータ収集の方法も，そして分析の方法も必然的に込み入ったものとなる。

しばしば，実験といえば処置群と対照群の結果を比較して処置の効果を検証する方法をすべて引っくるめて指す場合がある。しかし，厳密には無作為化を欠くものは実験としての要件を満たしておらず，準実験（疑似実験）と呼ばれる。上で紹介した不等価群処置前処置後計画は代表的な準実験であるが，これ以外にもより質の高いエビデンスが得られるよう様々な準実験のデザインが提案されている。

4. メタ分析

先に，コクラン共同計画がRCTをエビデンスの強さから臨床研究のトップに位置づけていると述べた。しかし，エビデンスは工夫次第でさらに強めることができる。その工夫とは，メタ分析（メタアナリシス）である。メタ分析とは，同じ研究仮説に関して実施された複数の研究について，得られた結果を統計的に統合する方法のことである[6]。コクラン共同計画では，RCTの結果をメタ分析して得られた知見が最も強いエビデンスを与えると規定している。以下，メタ分析の考え方を簡単に紹介しておく。

まず，統合の対象となるRCTがM個あり，それらの処置効果（処置後得点の平均値差）をD_mとする（$m=1,\ldots,M$）。簡単のために，各研究では処置群と対照群にそれぞれn_m人の患者が無作為に割り当てられており，個人のデータは共通の分散σ^2に従って変動するとする。このとき，各処置効果D_mを，真の効果Δと誤差E_mによって$D_m=\Delta+E_m$と表す。誤差E_mは各研究でどの患者がたまたま対象者として選ばれ各群に割り当てられたかによって偶然的に変動するもので，独立に正規分布に従うとする。これを$E_m \sim N\left(0, \sigma_m^2\right)$と書くことにすると，$\sigma_m^2=2\sigma^2/n_m$であることが導ける。すなわち，各群に割り当てられた患者数n_mが多いほど，処置効果D_mの変動は小さく安定しているということである。言いかえれば，$1/\sigma_m^2$がそのRCTのエビデンスの強さに相当することになる。とすると，真の処置効果Δを見積もるには，単にD_mの平均$\bar{D}_m\left(=\sum D_m/M\right)$を取るのではなく，強いエビデンスを与える研究により重みを持たせて結果を統合するほうがより合理的である。そこで，重み$w_m=1/\sigma_m^2$を考え，真の効果を$\hat{\Delta}=\dfrac{\sum_m w_m D_m}{\sum_m w_m}$という重み付き平均で見積もることを考える[7]。これがメタ分析の基本的な考え方であり，$\hat{\Delta}$がメタ分析により統合された処置効果である。

---

[6]「メタ」には「高次の」とか「超越した」といった意味がある。個々の研究で行われた分析（これを一次分析などと呼ぶ）をさらに分析することから，このような呼ばれ方をする。ちなみに，二次分析という言葉もあるが，これは別の目的で得られた過去のデータを再分析することで新たな知見を見出すことを指し，メタ分析とは少し意味が異なる。

[7] 実際には分散σ_m^2は未知であるので，これを推定値$\hat{\sigma}_m^2$で置き換えて重みを算出する。

メタ分析を適用した最初期の研究として，1977年に発表されたスミスとグラスの論文がある。この論文では，心理療法の効果を検証した375の研究結果を統合することで，心理療法が恐怖や不安の低減，自尊感情の向上に効果的であるという結論を導いている。学術研究の成果がデータベース化され，インターネットを通じて膨大な結果に簡単にアクセスできるようになった今，メタ分析はその威力をいかんなく発揮できるようになった。

5. エビデンスに基づく政策形成

コクラン共同計画に刺激され，2000年にキャンベル共同計画（Campbell Collaboration）が設立された。これは，広く司法，教育，社会福祉といった社会政策一般に関する提言を行う団体である。キャンベル共同計画は，公的政策は実験を通じてより良いものにしていく必要があるという考えに基づき，エビデンスにもとづく政策提言を積極的に発信している。コクラン共同計画と同様に，過去に実施された研究結果を収集して質を評価し，高いエビデンスを示すものを選りすぐって結果を統合することで政策の効果を評価するという方法がとられている。研究計画としてRCTを最も重視し，その結果をメタ分析したものが最も強いエビデンスとなると規定しているところもコクラン共同計画と同様である。キャンベル共同計画のウェブサイトでは，これら一連のメタ分析の結果がシステマティックレビューとして一般に公開されている[8]。

アメリカでは，2002年に「エビデンスに基づく教育」を実現するため連邦教育省にWhat Works Clearinghouse（WWC）というシンクタンクが設置された。WWCもRCTを重視し教育分野の研究をエビデンスの高さで格付けし，メタ分析によって統合された結果を報告書として公開している[9]。ここには，国や自治体単位で実施するようなものから学校や教師が自らの裁量で実施できるようなものまで様々な政策や実践が含まれている。一方，こうした事業を通じて，別の深刻な事態も明らかになってきた。それは，教育に関しては，エビデンスとして利用できるような質の高い研究自体がそもそもあまり存在しないということである。ある報告によれば，WWCが公表した500あまりの報告書のうち，WWCがエビデンスとして利用できるだけの基準をクリアしていると見なした研究（RCTと一部の準実験）を1つでも統合の対象にできた報告書は全体の3割にも満たないという[10]。実際，WWCの報告書には，ある教育実践が有効であるかないかが明らかになったというよりも，現時点ではどちらを支持する根拠も乏しくはっきりとしたことは言えないという論調のものが少なくない。

教育分野に限らず，エビデンスに基づく政策形成を実現するには確かな根拠として利用できる研究が幅広く行われることが必要である。このためには，政策づくりに関わる人々がエビデンスの価値に目を向けるだけでなく，その質がデータ収集と分析のやり方如何で大きく左右されることを知っていなくてはいけない。統計学の世界では，しばしば“Garbage in, garbage out.”（ゴミから生まれるのはゴミだけである）という格言が引き合いに出される。得られたデータの質が低いとどんなに凝った分析をしてもたいした知見は得られないという意味であるが，同じことはエビデンスに基づく政策形成にも当てはまるように思われる。わが国においても，産学官が連携することで真に政策形成に資するフィールド実験がもっと増えることを期待したい。現在に至るまで，政策形成のためと称して多くの調査研究が官主導で行われているが，実際にはその役目を果たさないのではないかと危惧している。

[8] https://www.campbellcollaboration.org/
[9] https://ies.ed.gov/ncee/wwc/
[10] Malouf, D. B., & Taymans, J. M. (2016). Anatomy of an evidence base. *Educational Researcher*, *45*(8), 454-459.