RIETI Policy Discussion Paper Series 26-P-014

EBPM（エビデンスに基づく政策形成）入門
第7話　信頼できるエビデンスの構築を目指して

関沢 洋一
経済産業研究所

RIETI
Research Institute of Economy, Trade & Industry, IAA
独立行政法人経済産業研究所
https://www.rieti.go.jp/jp/

RIETI Policy Discussion Paper Series 26-P-014
2026 年 6 月

EBPM(エビデンスに基づく政策形成)入門[1]

第７話　信頼できるエビデンスの構築を目指して

関沢洋一（独立行政法人経済産業研究所）

要　旨

・エビデンスを踏まえて政策立案が行われる EBPM が機能するためには、信頼できるエビデンスが構築されることが重要である。
・しかし、現実には、個々の政策を担当する行政機関、当該政策によって利益を受ける人々、そして研究者においても、政策介入の効果を水増しするインセンティブがあり、信頼できるエビデンスが構築されない場合がある（EBPM のハイジャック）。
・行政機関が行う効果検証はお手盛りとなる懸念があり、民間シンクタンクが行政機関の委託を受けて行う効果検証は発注側に忖度する懸念があり、中立的で信頼できる効果検証が行われていない可能性がある。
・EBPM において日本に先行しているアメリカの例なども参考にしつつ、信頼できるエビデンスを構築するための環境づくりを行うことが必要である。

キーワード：PBEM（政策に基づくエビデンス形成）、EBPM のハイジャック、効果の水増し、システマティックレビュー
JEL classification: H11 H43

[1]本稿の原案は、経済産業研究所（RIETI）のポリシー・ディスカッション・ペーパー検討会で発表を行ったものである。

第７話　信頼できるエビデンスの構築を目指して

（要約）

- エビデンスを踏まえて政策立案が行われる EBPM が機能するためには、信頼できるエビデンスが構築されることが重要である。
- しかし、現実には、個々の政策を担当する行政機関、当該政策によって利益を受ける人々、そして研究者においても、政策介入の効果を水増しするインセンティブがあり、信頼できるエビデンスが構築されない場合がある（EBPM のハイジャック）。
- 行政機関が行う効果検証はお手盛りとなる懸念があり、民間シンクタンクが行政機関の委託を受けて行う効果検証は発注側に忖度する懸念があり、中立的で信頼できる効果検証が行われていない可能性がある。
- EBPM において日本に先行しているアメリカの例なども参考にしつつ、信頼できるエビデンスを構築するための環境づくりを行うことが必要である。

EBPM と紛らわしい言葉として PBEM があります。Policy-based evidence making の略語で、日本語に直訳すると「政策に基づくエビデンス形成」になります[1]。図７－１が EBPM と PBEM について端的に示したものになります。

EBPM においては信頼できるエビデンスが中立的・客観的に作られて、政策形成に当たってそのエビデンスが参照されることが目指されます。PBEM では、反対に、政策に関係する者がその都合に合わせてエビデンスを作ることが目指されます。PBEM で作られるエビデンスは信頼しにくいものとなります。

図７－１　EBPM と PBEM のイメージ

PBEM という言葉は比較的新しいものですが、PBEM らしきものは長きに

[1] List (2024)では Policy-Based Evidence という表現が使われていますが、ここでの議論とは異なります。

渡って日本の政治と行政においてほとんどその問題を疑われることなく推進されてきました。日本の行政機関によく見られることとして、自分たちが行っていることや実現したいことを正当化する上で役に立つ情報・エビデンスを見つけてきたり自分たちで作ったりして、それらを予算要求や審議会での審議などで示して関係者を誘導し、自分たちが行っていることや実現したいことをサポートしない情報やエビデンスを無視してきました。これが PBEM の典型例になります。

似たことは役所に限らず、政界でも民間企業でも司法界でも学会でも広く見られることで、絶対に悪いとは言いきれません。

しかし、EBPM や EBM（エビデンスに基づく医療）といったエビデンスに基づくアプローチではこのようなことは推奨されません。信頼できるエビデンスを構築して、構築されたエビデンスは隠すことなく多くの人々がアクセスできるようにして、それらのエビデンスを基礎として意思決定を行うことがエビデンスに基づくアプローチの本来のあり方です。政策を立案する側の都合によってエビデンスが歪められるのは避けるべきことになります。

本稿では第 1 節において、エビデンスに基づくアプローチを象徴的に示すものとしてシステマティックレビューを取り上げます。第 2 節では EBM と EBPM のハイジャック問題と効果の水増しについて取り上げます。第 3 節では政策介入の効果検証を行う主体がエビデンスの信頼性に及ぼす影響について、民間企業が効果検証を行う場合を中心に取り上げます。第 4 節において日本における今後の方向性について取り上げます。

第 1 節　システマティックレビュー

第 2 話と第 3 話において、政策介入の効果を明らかにするための代表的な手法としてランダム化比較試験（RCT)、回帰不連続デザイン（RDD)、差の差分析（DID）を取り上げました。これらの手法は適切に運用される限りは信頼できるエビデンスを提供しますが、それぞれの効果検証が示すエビデンスが信頼できるものであっても、似たような介入の効果検証が複数回行われれば、示される効果にバラツキが生じます。システマティックレビューはこのようなバラツキに対処して介入についての効果を総合的に判断しようとします。

より具体的には、システマティックレビューでは、健康診断が重大疾患や寿命の延びに及ぼす影響など（第 5 話参照)、特定のテーマについて学術研究用の検索サイト（PubMed, Web of Science, Scopus など）の検索機能を使ったり過去の研究で引用されている文献を探したりするなどして、世の中にある関連研究を網羅的に探します。見つかった研究のうち質の低いものを除いて、メタ

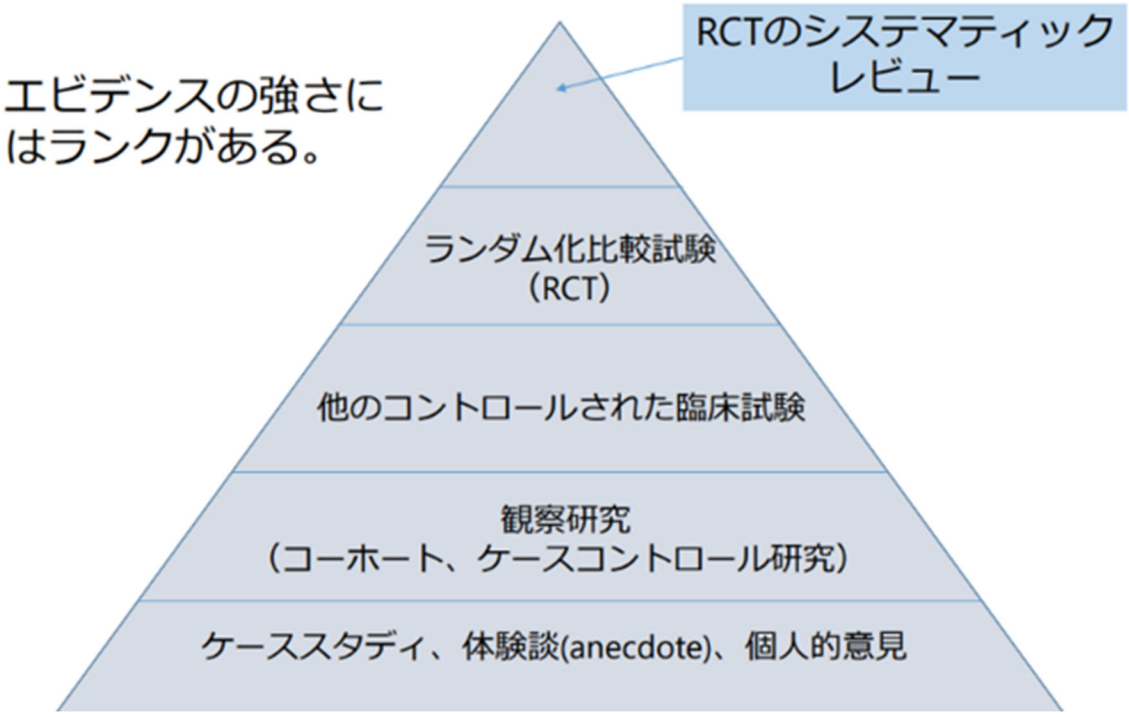
アナリシスと呼ばれる手法によって全体としての効果を数値化して示します[2]。

このように対象となる研究を探し出して質によって絞り込む際に恣意的な運用が入り込む可能性もあるため、実際にレビューを行う前に、レビューの手順（プロトコル）を事前登録することが多くなっており（決められたウェブサイトで公表します）、後出しじゃんけん的な運用を避けることが目指されています[3]。

システマティックレビューが多く作られているのは医療で、EBM において重要な役割を果たしています。EBM を主導する NGO にコクランがあり、社会政策に関するシステマティックレビューを作っている代表的な NGO としてキャンベル共同計画があります。

システマティックレビューの中でも特に RCT のシステマティックレビューは最も高度なエビデンスを示すものとされています（図７－２）。

図７－２　エビデンスのピラミッド



（出所）Greenhalgh (2014), P18 より作成。

システマティックレビューの例（スケアドストレート）

コクランとキャンベル共同計画の両方に掲載されたシステマティックレビュ

[2] メタアナリシスがないシステマティックレビューがあり、システマティックレビューでないメタアナリシスもあります。両者は同義ではありません。

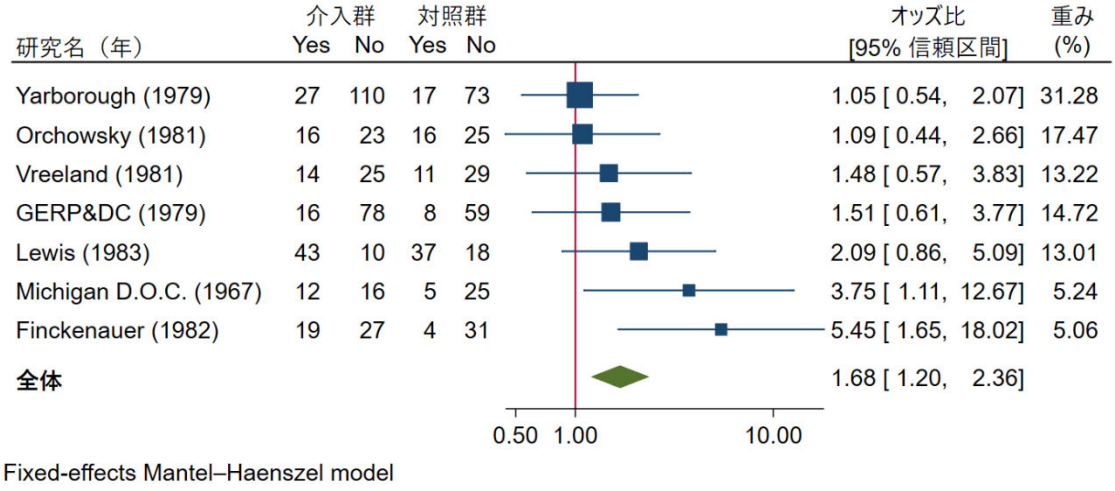
[3] Pieper and Rombey (2022)

ーを紹介します[4]。スケアドストレート（scared straight）という少年の非行予防のためのプログラムのシステマティックレビューです。スケアドストレートは、非行に走るリスクが高い未成年者を対象として、彼らを刑務所に連れて行って受刑者と交流させることによって、未成年者の犯罪や非行を予防しようとするものです。

このレビューの著者は、学術誌の検索サイトを使った検索、関係しそうな研究を行っている研究者への問い合わせなど、様々な手段を使ってスケアドストレートの効果検証を行った社会実験の結果を示した研究や報告書を網羅的に探していきました。個々の州で行われた実験は必ずしも論文化されていないのですが、これらも探そうとしました。見つけた文献について、ランダム化が適切に行われたかどうか、アウトカムが計測されているかをチェックして、最終的に７つの RCT が残りました。

それぞれについてオッズ比[5]を計算して、メタアナリシスによって全体としての効果を算出しました。図７－３がシステマティックレビューにおけるメタアナリシスの結果を示すもので、フォレストプロットと呼ばれます。

図７－３　スケアドストレートのメタアナリシス



（出所）Petrosino et al. (2013b)の Analysis 1.1.に示された 7 つの研究のイベント数を用いて統計ソフトの STATA を使って Analysis 1.1.のフォレストプロットを再現した。

[4] Petrosino et al. (2013a); Petrosino et al. (2013b)

[5] オッズ比は１を超えれば、アウトカムとなる二値変数（この場合は犯罪を行った場合が１、行わなければ０）の発生割合が増え、１を下回れば減ると判断され、９５％信頼区間が１をまたがなければ有意差があると判断されます。オッズ比が１を下回ると犯罪予防につながることになります。

図７－３の左上に研究名（年）とあり、たとえばYarborough (1979)とあるのはYarboroughらの1979年の研究となります。Yesとあるのはイベント（この場合は犯罪や非行）を発生させた人数、Noとあるのはイベントを発生させなかった人数で、介入群と対照群のそれぞれについて記載されています。個々の介入のオッズ比はこの数値だけから計算できます。Yarborough (1979)の場合、オッズ比は1.05[0.54,2.07]となっていて、これはオッズ比の点推定値が1.05で95%信頼区間の下限と上限がそれぞれ0.54と2.07であることを意味します。図７－３の真ん中にこれを図示したエラーバーが記載されています。95%信頼区間が１をまたがっていますので、Yarborough (1979)だけを見れば有意な効果はありません。他の６つの研究の結果も同様に示されています。一番下の２つの研究（Michigan D.O.C.とFinckenauer）は95%信頼区間が１をまたがず、点推定値は１を超えているので、いずれも負の効果があった（犯罪を増やした）ことになります。

図７－３の下の方にある「全体」がメタアナリシスの結果としての全体の効果を示します。オッズ比の点推定値が1.68で１を上回っており、95%信頼区間が1.2～2.36で１を含まないので、スケアドストレートは有意に犯罪を増やした（負の効果があった）という分析結果になります。◆が全体としての効果を図示しています。

仮にシステマティックレビューがなければ、スケアドストレートに反対する人は７つの中でオッズ比が一番高い研究であるFinckenauer(1982)を使って自分の主張を正当化し、スケアドストレートに賛成する人はオッズ比の一番低いYarborough(1979)を使って有意差がないから有害とは言えないと主張しそうです。ただしオッズ比が１を下回る研究はないため、スケアドストレートに犯罪予防効果があると主張できる研究はありません。

仮の話ですが、悪い結果は公表したくないという動機から、Michigan D.O.C.やFinckenauerの効果検証を実施した研究者が分析結果を闇に葬ったり、このシステマティックレビューの著者がこれらの分析結果をシステマティックレビューの対象から外したりしたらどうなっていたでしょうか。

その場合、この２つの研究はこのメタアナリシスに含まれていなかったことになります。実際に統計ソフトを使って推計したところ、この２つの研究を外すと、オッズ比は1.35（95%信頼区間は0.93～1.97）となり、効果があるとは言えないものの、有意に害があったとも言えないという間違ったメッセージを送ることになったはずです。

第２節　EBMとEBPMのハイジャック問題と効果の水増し

前節から示唆されるとおり、システマティックレビューはいくつかの課題を

抱えています。１つめは、効果検証研究の抽出段階で恣意的な抽出が行われるリスクです。２つめは、システマティックレビューの対象となる個々の効果検証研究において効果が操作されるリスクです。３つめは、効果検証の結果が公表されないためにシステマティックレビューの対象にすることができないことです。このうち、１つめは上記の事前登録である程度対応できますが、２つめと３つめは外部からは把握しにくい面があります。

たとえば、製薬会社は、収益を増やすために、自社が開発した医薬品に高い効果があるように見せたいというインセンティブがあり、操作によって自社が開発した製品の効果を高くする分析結果を出そうとしたり、効果がないとか乏しいという分析結果を意図的に公表しないようにしたりすることが過去にありました。

EBM に関しては、スタンフォード大学の Ioannidis 教授が、利害関係者（効果がないというエビデンスが明らかになると困る人々が中心）によって EBM がハイジャックされている（乗っ取られている）と指摘しました[6]。私も日本の EBPM において似たようなことが起きていることに気づいて、EBPM のハイジャックという言葉を使ったことがあります[7]。

EBM や EBPM のハイジャックは、基本的には、効果がなさそうなものを効果があるように見せることも含めて、介入の効果を上振れさせようとすることが多いので、以下では効果の水増しという言葉を使います。

誰がどのような理由で効果の水増しを望むのか

まず、誰がどのような理由で効果の水増しを望んでいるのかを明らかにし、実際にどのような手段を使って効果の水増しが引き起こされるのかを述べます。

①ビジネス上の利害関係者

一番わかりやすいのは製薬会社で、自社が開発した医薬品に効果がなければ膨大な開発費が無駄になるので、医薬品の効果検証である治験において効果が高く出ること（あるいは、効果がないという結果の出た治験が公表されないようにすること）を望みます。医療関係の様々な機器、日本で言えばトクホや機能性食品でも似たようなインセンティブが開発企業にはあります。

また、定期健康診断、がん診断、ストレスチェックなどの医療関連サービスはビジネスとして確立していて多くの雇用も創出されているので、これらの業

[6] Ioannidis (2016)
[7] 関沢 (2022)

務に携わる人々や企業においては効果があることを求める（効果がないというエビデンスが示されたら困る）場合が多いものと思われます。

医療・健康分野に限らず、たとえば、企業を対象とした補助金では補助金事業の実務を広告会社などが担っています。中小企業相手の補助金では申請をサポートするために多くの経営コンサルタントが関わっています。こうした人々は、企業への補助金に効果がないというエビデンスが示されてその補助金が廃止されることになれば、悪影響を被ることになるかもしれません。

②公務員や政治家などその政策に公的部門で関わる人々

効果がないという分析結果が出ることを当該政策の担当部局の人々が嫌がる場合はしばしば見られます。個々の政策介入に効果がないことが判明すれば、その政策の実施を決めた役所の幹部や政治家の顔に泥を塗るかもしれませんし、無駄なことをしたという非難を野党の政治家やマスコミから受けるかもしれませんし、極端な場合には、自分のいる組織のお取り潰しが取り沙汰されるかもしれません。

このことはアメリカなどプログラム評価が盛んに行われてきた国では古くから知られており、プログラム評価への政治的圧力としてしばしば取り上げられてきました[8]。日本でも評価結果を歪めるお手盛り型の評価行動として行政官が行う「作為的な評価行動」の研究があり、政策評価実務の従事者の評価行動が評価結果を歪める方向で働くことが指摘されてます[9]。

③研究者

政策介入に効果がないという分析結果が出ると困る事情は分析に携わる研究者にもあります。研究者は分析結果を論文化して学術誌に投稿します。学術誌側は投稿された論文案をチェックして（査読と呼ばれます）[10]、掲載に値する質の高い論文と判断されたもののみが最終的に学術誌に掲載されます。政策研究に限らないのですが、効果があるという結果を示さない研究を掲載したがらない（統計学的に有意でない結果が出る研究を評価しない）傾向が学術誌側にあることが指摘されています[11]。

---

[8] Fels (2022); Pleger and Sager (2018); Weiss (1993)

[9] 西出 (2020, 2023)

[10] 査読では、通常は２～３名の研究者（レフリー、レビューアー）にその論文案についてコメントしてもらい、そのコメントが適切に反映されていると編集長と全てのレビューアーが同意した場合に限って、学術誌に掲載されます。学術誌への掲載は研究者の評価に大きな影響を及ぼすので、多くの研究者は一流の査読付き学術誌への掲載を目指して研究活動を行います。査読については植原 (2025)に詳しく述べられています。

[11] Abadie (2020); Chopra et al. (2023)

こうした事情があると、論文書きを本業とする多くの研究者としては介入に効果がないという分析結果を出すことにためらいを感じることになります。このため、効果があるという結果が出やすい分析を行おうとしたり（後述します)、効果がなさそうだという結果が出そうな研究の実施を最初から避けようとしたりするインセンティブが研究者に生じることになります。

効果の水増しを起こすための手段

本当は効果がない介入について効果があるように見せるための手段として主要なものは２つあります。１つめは分析を工夫することによる対応で、２つめは効果がないという結果が出たものを公表しないことです。もう１つめったにはないとは思いますが、データそのものを改ざんする場合が過去に見られました。ここでは EBPM に先行する EBM の例も含めて紹介します。

①後出しじゃんけんで効果があるという分析を意図的に残す（p ハッキング）

効果検証に関する分析を行っていると、複数の分析手法のどれを使うか（RDD か傾向スコアマッチングなど)、傾向スコアマッチングなどでどの変数を調整に用いるか、欠損値（アンケート調査などで一部の質問が回答されていない場合などに生じる数値の空白）や外れ値（一部の者に他の大部分から著しく乖離した数値が計測された場合のその値のこと）をどのように取り扱うかなど（外れ値を分析から外すかどうか、欠損値のある研究対象者を丸ごと分析から外すか、欠損値について何らかの数値を代入するかなど)、分析の方針によって分析結果が変わることがしばしばあります。

データにアクセスして分析結果を見てから分析方法を変えるのは望ましくないこととされているのですが[12]、効果があるという結果を出したい場合、いろいろなパターンで分析してみて結果を見てからまた別の分析をして、有意に効果がある（あるいは一番大きな効果が見られる）という結果が出たものだけ残すということが実際には可能です。

研究領域や投稿する学術誌のレベルによって違うとは思いますが、何度分析を繰り返したかを報告することは求められていない場合が多いので、特に有意か否かがギリギリのところにある場合には、分析結果を見た後で分析手法等を変えることによって、有意かどうかという肝心な結果をコントロールすることが可能となります。これは p ハッキングと呼ばれます[13]。p ハッキングの p は統計学で用いられる p 値のことです（第 2 話参照)。似たような言葉として、

[12] Nosek et al. (2018)
[13] Brodeur et al. (2020)

いいとこ取り的に効果があった部分だけ選択的に公表したりすることはチェリーピッキングと呼ばれます[14]。

②良い結果だけ公表して効果のない結果は公表しない（公表バイアス）

データを分析して出てきた結果の中には効果があったというものもあれば、効果があるとは言えないというものもあります。これらのうち、効果があったという検証結果だけは公表し、効果があるとは言えないとする検証結果は公表しない場合が現実にあります。この点は医療の世界で特に問題になっており、医学的介入の効果を検証する研究を行って効果があるとは言えないという結果になった場合には医学雑誌に掲載されない傾向があります[15]。効果があったという分析結果のみが公表され、効果があるとは言えない、または望ましくない結果は公表されないことを公表バイアス（publication bias）と呼びます。

③データを改ざんして効果があるように見せる

効果の水増しを実現するためにデータそのものを改ざんすることがあります。ここでは医療関係で日本において起きた事件を紹介します。ディオバンという降圧剤が他の降圧剤よりも循環器疾患の予防効果が大きいかどうかを検証する RCT が、ディオバンの製造・販売を行う製薬会社の支援の下、日本の大学で行われました。この研究では、ディオバンを投与しなかった群の循環器疾患の発生件数が水増しされたために、ディオバンが他の降圧剤よりも効果があるという間違った結果が報告されました（後に撤回されました）[16]。

効果の水増しはどのような問題を引き起こすのか

本当は効果がない政策介入が効果ありとされる場合、その政策介入が継続されることにつながり、意味のないことにリソースが使われることになります。

政策介入とは少し異なりますが、介入の効果の水増しが特に問題になっているのは医療における公表バイアスで、RCT のシステマティックレビューにおいて問題になっています。特定のテーマに関する全ての研究がシステマティックレビューの対象となれば問題ないのですが、効果があることを示せなかった研究が対象から外れると、メタアナリシスの結果として明らかにされる効果が本当の効果よりも上振れすることになり、エビデンスを利用する側に誤解を生じさせることになります[17]。たちの悪いことに、外部の人間には公表バイアスが

[14] Tsang (2025)
[15] Ioannidis et al. (2014)
[16] Sawada et al. (2009)
[17] McGauran et al. (2010)

あるかないかがわからないため、システマティックレビューによって明らかになったエビデンスを信じていいか疑心暗鬼になります。

医療の世界で有名な例として、抗うつ薬の SSRI の効果についての公表バイアスがあります[18]。Kirsh らは、アメリカ食品医薬品局（FDA）への情報公開請求によって公表されていなかった臨床試験の情報を入手して、SSRI の効果についてのシステマティックレビューを行いました。その結果、従来のシステマティックレビューで示されていたよりも SSRI の効果が小さいことが明らかにされました。他の有名な例として、インフルエンザの薬であるタミフルの効果について公表バイアスによる効果の水増しが問題になりました[19]。

近年は、いろいろな分野でシステマティックレビューが盛んに作られるようになっており、公表バイアスを巡る問題が更に深刻になっている可能性があります。最近の例では、ナッジ[20]について公表バイアスによって効果の水増しが生じているのではないかという指摘が出ていて、論争になっています[21]。

第３節　効果検証の主体毎に見たエビデンスの信頼性を巡る課題

政策介入の効果検証を行う主体は行政機関の職員、大学などの研究者、民間企業の３つに分かれます。それぞれについての課題を見ていきます。

行政機関の職員が行う効果検証

政策評価法による政策介入の効果検証は行政機関内部における自己評価を原則としています。行政機関ではありませんが、IT 企業では A/B テストを始めとして自社の取組がもたらす効果を社内で把握しています。効果のない取組を行えば業績の悪化につながるという背景事情があります。

理想的には、行政機関内でも IT 企業のように効果検証を自ら行って政策改善につなげるという流れを作れればいいのですが、いろいろな事情があって、なかなかうまくいきません。

このような事情をより具体的に挙げると、政策介入の効果検証の困難さを生じさせる一般的な事情として、①政策介入の多くはある程度の長期間をおいて効果が発現することが想定され、A/B テストのように短期間で結果が分かる場合が少ないこと、②分析に必要なデータを取得ができない場合が多いこと（第

[18] Kirsch (2014)
[19] Jefferson et al. (2014)
[20] ナッジとは Thaler and Sunstein (2008)で提唱された発想で、人々を望ましい行動に導く上で、強制的な手法を用いずに自発的に望ましい方向に向かうように仕向けるための介入を指します。
[21] DellaVigna and Linos (2022); Maier et al. (2022); Mertens et al. (2022)

６話参照)、③後々で効果検証を行えるような制度設計をしてない場合が多いこと（第４話参照）があります。これらの一般的な事情に加えて、行政機関内部で行う効果検証の場合に固有の事情として、④政策介入を効果的に行っているように見せようとする「お手盛り」へのインセンティブが行政機関にはあること、⑤行政機関内部で効果検証の分析に携わる専門家が少ないこと、⑥効果検証への取組は時間がかかることが多く、短期的な対応を求められる行政機関が自ら行う業務にしにくいことがあります。

以上の事情により、政策介入の効果検証は行政機関が自ら行うよりも、専門的知識と独立性をもった外部の第三者に時間をかけて行ってもらう方が望ましいという考え方がでてきます。

研究者が行う効果検証

経済学者をはじめとして、因果推論に対する専門的な知識を有する研究者が政策介入の効果検証に携わる事例は多々見られます。

研究者側のインセンティブは、行政機関が保有する業務情報にアクセスすることによって研究を行い、論文が書けることです。研究者の評価のかなりの部分は、レベルの高い査読付き学術誌に自分の論文が掲載されることによって決まります。最近の論文はデータ（特に一般に公開されていない政府統計や業務情報）がないと書けないものが多く、行政機関の業務情報にアクセスできるかどうかは研究者にとって死活問題になりかねないくらい重要になります。

その一方で、業務情報を提供する側である行政機関としては、研究者が業績を上げることにおつきあいするインセンティブはあまりありません。逆に、研究者に業務情報を提供することによって望ましくない分析結果が公表されることは避けたいというインセンティブがあります。

EBPM の推進や学術研究の水準の向上を目指すためには、研究者による業務情報のアクセスを増やすことが望ましいです。上述した p ハッキングやチェリーピッキングのように、信頼できないエビデンスを構築するインセンティブは研究者においても存在するのですが、その一方で、研究倫理として正直さや誠実さ、真理の探究が研究者には求められており[22]、ある程度の倫理意識を多くの研究者は有しているように思われます。また、研究論文は査読によって第三者からのチェックが入るため、査読誌（特に一流誌）に掲載された政策介入の効果検証は信頼度が高くなります。

ただ、現状では、研究者が業務情報にアクセスするのは研究者と行政機関の

[22] 日本学術会議「声明　科学者の行動規範―改訂版―」平成２５年１月。
https://www.scj.go.jp/ja/scj/kihan/kihan.pamflet_ja.pdf

個別のやり取りの中で決まるのが通常です。このような個別のやりとりでは、行政機関にとって有意義な情報が得られることを研究者は強調する一方で、行政側には迷惑をかけないと事前に伝えるなどして行政側を説得している場合もあるかもしれません。このことは、政策介入の効果の水増しや、行政側にとって望ましくない研究成果を公表しないことにつながるかもしれません。仮にこのようなことが起きていれば、研究倫理との関係で微妙な問題を生じさせますし、EBPM という観点から見れば望ましくないです。

このような事態にどう対応すべきかについてはあまりいい答えがないのですが、１つの方向としては業務情報の提供を仕組みとして確立することが考えられます。第６話で取り上げたいくつかの事例（デンマーク統計局、アメリカのセンサス局・内国歳入庁・NVSS［National Vital Statistics System］、日本の財務省による一定の条件の下での研究者への業務情報の提供など）は、業務情報の提供の仕組みが成文化されており、行政機関にとって都合が悪い情報は出さないという懸念は減らせるのではないかと思われます。

民間企業が行う政策介入の効果検証（日本の場合）

政策介入の効果検証を行う主体として大きな役割を負っているのは、民間シンクタンクやコンサルティングファームと呼ばれる民間企業です[23]。

これらの企業は行政機関からの委託を受けて効果検証を行います。行政機関の側では、当該政策の企画及び実施に携わる部局（その政策を推進するスタンスに立つのが普通）が委託関連の業務を担う場合が多いです。政策介入の効果検証によって効果がありそうにないという分析結果が出ることがしばしばありますが、これは多くの担当部局としては見たくないものです。

そうすると、効果検証を受託した民間企業に対して、悪い分析結果の非公表も含めた効果の水増しを求める圧力が行政機関からかかるリスクが生じます。また、明確な圧力はなくても、受託者の側において行政機関の意図をくみとって効果を水増ししようとするインセンティブ（いわゆる忖度）が生じます。

学術研究者と異なって、民間シンクタンクの多くは顧客の意向に可能な限り応えていくことが重要な使命なので、今の仕組みの下では忖度に対する歯止めは弱いです。この点は、たとえば、顧客からの独立性が法律上明記されている公認会計士とは異なっています。

その結果、不正とまでは言えない範囲で、効果検証を受託した民間企業が p ハッキングやチェリーピッキングを行う余地が出てきます。たとえば、恣意的な分析を行いやすい傾向スコアマッチングを主たる分析方法として選択して、

[23] 伊芸 (2024)

更に傾向スコアマッチングにおける調整変数の選択の仕方も工夫して（第３～４話参照）、あたかも政策介入に効果があるかのような分析結果を何とかひねり出すといったことです。

仮に、発注者側への配慮に起因した効果の水増しが民間シンクタンクによって頻繁に行われているとすれば、エビデンスの信頼性が揺らぎ、EBPM が機能不全に陥ります。

忖度以外にも、民間シンクタンクが行う評価の中には首をかしげざるをえないものがしばしば紛れ込んでいます。先行研究のチェックが十分でない（日本語で書かれてネット上に載せられた文献が中心で諸外国の学術研究への依拠が少ない）、分析手順の説明が報告書に十分に記載されずどうやって分析したかがわからない、分析時間を切り詰めるなど、研究者があまり行いそうにない問題が見受けられます。

その背景としては、①分析に携わる者の多くが因果推論を十分に学び最新の研究動向もフォローしている研究者ではなく、プロの仕事とは言いがたい、②政府の予算が１年単位なので納期について時間の制約を受ける、③委託元の行政機関内に評価報告をチェックできる専門家が少ない、④報告内容を有識者がチェックする機会を設ける場合はあるものの、原データにアクセスすることなく短時間の会合でコメントをもらうだけで、実際にはお墨付きを与える場になりやすい、といった事情があるように思います。

信頼に値するエビデンスを得るためには今のやり方を抜本的に見直す必要があります。その際に参考になりそうなのは以下で述べるアメリカの評価企業と次節で述べるアメリカ政府の文書（OMB 文書）です。

アメリカの評価企業

アメリカではプログラム評価を行う組織が多く存在します[24]。実験大国のアメリカでは EBPM という言葉が登場する以前から、政策プログラムについて RCT による効果検証が行われてきましたが、これらの多くは大学のような研究機関ではなく企業によって担われています。評価に携わる企業（以下では「評価企業」）、特にエリート評価企業とでも呼ぶべき組織（MDRC、RTI、Mathematica、Abt Associates など）においては、大規模な RCT を実施したり高度な分析を行ったりすることが可能となっています[25]。評価企業という言葉を使ってはいますが、非営利組織（例えば MDRC や RTI）や従業員が保有する企業（例えば Mathematica）など、営利性は高くない場合が多いです。

[24] Peck (2018)
[25] Rossi (2003)

プログラム評価において委託元である行政機関から圧力がかかる場合があることはいくつもの研究で報告されています[26]。たとえば、プログラムに効果がないという結果が出た場合を中心としてその結果を公表しないようにする、あるいは、意思決定に影響を与えないようにするために公表のタイミングを意図的に遅らせる場合があることも指摘されています。

ただ、現実にどの程度アメリカのプログラム評価で効果の水増しが起きているかというと、印象に過ぎないのですが、意外に少ないように感じます。特に、日本における行政機関から民間企業への委託と比べると忖度は少ないように感じます。そう感じる背景として、多くの社会実験ではプログラムに効果がないことが示され、効果があったものでも限定的でした[27]。効果の水増しらしきものが見えないです。

アメリカのプログラム評価で忖度が起こりにくく、評価企業の独立性やモラルは高いのではないかと推測される背景事情があります。

第一に、クライアントに忖度することが知られればその評価機関の評判が落ちるので、ブランドを守るためには忖度はせずに正直に作る面があります。この意味では評価企業はシンクタンクというよりも監査法人に近いところがあります[28]。例えば、MDRC のウェブサイトでは独立性の重要性が明記されています[29]。

第二に、全米評価学会（AEA, American Evaluation Association）などの評価者の団体では倫理指針があり、そこでは独立性や中立性の重要性が記載されているため、ここから大きく乖離するのは難しいと推測されます[30]。

第三に、評価企業の中には民間企業にはなっておらず公益性を標榜しているものがあります。

最後に、後述するとおり、政府の文書においても評価者が科学的な知見を重視することや独立性や客観性の重視を謳っており、発注者側にも制約がかかります[31]。

ただ、プログラム評価も玉石混合で、上述したようなエリート評価企業が行った本格的な評価のように質の高いものもあれば、アカウンタビリティ対策などで小企業などが委託を受けて行った形ばかりの評価では信頼度の低いものも多いようです[32]。

---

[26] Brown and Klerman (2012); Morris and Clark (2013)
[27] Besharov (2009); Thomas (2021)
[28] Brown and Klerman (2012)
[29] https://www.mdrc.org/about
[30] American Evaluation Association (2018); Schmidli et al. (2023)
[31] US Office of Management Budget (2020)
[32] Rossi (2003)

第４節　日本における今後の方向性

アメリカの行政管理予算局（OMB）が発出した文書

日本の EBPM はまだ始まったばかりで、ノウハウはまだあまり蓄積されていません。数十年にわたって試行錯誤し続けてきたアメリカから学べるものは多いです。

信頼のあるエビデンスを構築していく上で参考になりそうなものとして、アメリカの Evidence Act（Foundations for Evidence-Based Policymaking Act of 2018）に基づいて行政管理予算局（OMB）が発出した文書（以下では「OMB 文書」）があります[33]。ここには連邦政府のプログラムの評価の標準と実践が書いてあります（表６－１）。これを参考にしながら考えてみたいと思います。

表６－１　アメリカ連邦政府のプログラムの評価の基準とプラクティス

連邦政府のプログラムの評価の基準	
関連性と実用性	評価は行政機関の重要な問題に取り組み、ステークホルダーにとって有意義な情報源とならなければならない。
厳密さ	評価は、限界を明らかにしつつ、行政機関とその利害関係者が信頼できる正確な結果を生み出さなければならない。
独立性と客観性	評価は独立して客観的に行われなければならない。これらの原則は評価者の独立性と客観性によって決まる。
透明性	評価は、計画・実施・報告の各段階で透明性を保ち、説明責任を果たし、特定の結論に誘導されないようにする必要がある。
倫理	公衆を守り、その信頼を維持するため、評価は最高水準の倫理基準をもって行われなければならない。
評価のプラクティス	
１．評価のキャパシティを構築し維持する	
２．専門家への相談を効果的に行う	
３．各行政機関の評価ポリシーを確立し、履行し、広く共有する	
４．評価の設計と方法はあらかじめ特定する	
５．鍵となるステークホルダーを意味のある形で関与させる	
６．情報共有を戦略的に計画する	
７．参加者の倫理的な扱いを確実にする手段を講じる	
８．評価のためのデータマネージメントを促すとともに適切な利用を図る	
９．評価データを二次利用できるようにする	
10．独立性と客観性を守るためのポリシーと手続きを確立して維持する	

[33] US Office of Management Budget (2020)

（出所）US Office of Management Budget (2020)より筆者が作成。

OMB 文書を参考にした政策評価の基本的なあり方

OMB 文書はよく書かれているので、興味ある方は是非お読みください。以下は OMB 文書を踏まえつつ、効果検証を伴う評価について私なりに考える理想像を書きます。以下では OMB 文書で使われている「評価」という言葉を使っていますが、「効果検証」と同義です。

①評価の独立性と客観性

信頼できる評価を行うためには、評価者が発注元からの独立性を担保し、発注者に忖度しないことが必要になります。評価者は発注者側から必要な情報を入手するだけでなく、意見交換も積極的に行うべきですが、プログラムの効果の有無や効果の大きさの評価については、信頼される分析手法とデータに基づいて明らかにするものであり、発注者側の要請に基づいて操作することは許容されないことを明確にする必要があります。

②評価の透明性

発注者側にとって望ましい結果が出れば公表し、望ましくない結果がでれば公表しないというのは透明性を欠きます。作成された評価報告の扱いをどうするかはあらかじめ定めておく必要があります[34]。公表が好ましくなければ、評価開始前にその旨を発注者と評価者の間で合意する必要があります。国民に対する説明責任を果たす、公表バイアスを避けるといったこともあり、公表しない場合は限定的にすべきものです。報告書が情報公開請求にかかる可能性もあるので、報告書を公表しないで済ませられる場合は実際には少なくなるはずです。

OMB 文書は更に突っ込んでいて、評価者が評価を開始する前に設計と分析手法をウェブサイトなどで明らかにすることを求めています。おそらく、アウトカムを後から発注者側の都合のいいように変えたり、分析手法を途中で変えて p ハッキングのような後出しじゃんけんを行ったりすることを避ける趣旨のように思われます。

RCT のような実験型の効果検証やシステマティックレビューの場合はアウトカムだけでなく分析手法をあらかじめ決めておいて事前登録することは頻繁に行われます[35]。これに対して、介入を伴わずに既に存在するデータにアクセス

[34] O'Hara (2020)
[35] 長谷川他 (2021)

して実施する観察研究の場合には、データを見てから初めて元の分析プランの問題点に気づくこともあるため、分析手法を事前に決めるという方針を貫くのは厳しすぎる気もします。このような場合に、当初の分析プランと、事後的に決めた分析プランを分けて、明示的にその旨を論文に記載することが研究の世界では推奨されていますが（事後解析、post-hoc analysis）、政策介入の評価一般においても有用だと思います。

③業務情報へのアクセスの拡大（第 6 話参照）

評価者が評価に必要な情報にアクセスできなければ、適切な評価は行えません。これらの情報の多くは公表されていませんが日本政府や地方自治体の中に既にあります。守秘義務を課すなどの必要な措置を講じつつ、評価者に業務情報へのアクセスを認めることが重要になります。

④第三者によるデータの再利用

評価に用いられたデータの利用を評価者以外の研究者が求めた場合にそれを認めることは、分析結果の再現可能性を担保して、評価の妥当性を高める上で重要です。第三者が同じデータにアクセスできれば、元の研究の分析が本当に正しい結果にたどり着いているのかが検証できますし、そういう仕組みがあるだけでも、分析者が p ハッキングやチェリーピッキングを行うことを防ぐことにつながるかもしれません。

一例として、アメリカで行われている幼児教育のプログラム（ヘッドスタート）があります。このプログラムの効果検証として複数の RCT が評価企業によって行われていて、600 ページに及ぶ最終レポートが 2010 年に出されています[36]。この評価に際して収集されたデータは研究者がアクセスできるようになっており、実際にもいくつかの研究が行われています。

センシティブな情報の場合、第三者によるデータの再利用を認めることは難しい場合はあるかもしれません。そのような場合であっても、査読付き学術誌の中には分析のために作成したコードを提出させて、コードだけは公表する場合があります。コードがわかるだけでも不適切な分析が行われていないかどうかはある程度わかるはずで、特に分析プランがあらかじめ明らかにされている場合には、事前の予定との食い違いがわかるため、問題ある分析が行われていないかどうかを把握できます。

終わりに

[36] U.S. Department of Health and Human Services (2010)

本稿の目指したことは、読者に EBPM の影の世界を知っていただくとともに、政策介入の効果検証が歪むことを防ぐための取組を紹介することでした。こういうことを多くの方々に知っていただくだけでも忖度や p ハッキングによる怪しげな効果検証が減るのではないかと期待しています。

重要なことは、エビデンスは信頼の置けるものであるべきという見方が多くの人々に共有されることです。エビデンスはでっち上げでもかまわないとみんなが考えるようになれば、モラルは消滅して本当にでっちあげ的なエビデンスもどきがはびこることになります。それは EBPM の死につながります。

極めて当たり前のことだと思いますが、信頼できるエビデンスが構築されることは EBPM が成立する大前提です。

引用文献

Abadie, A. (2020). "Statistical Nonsignificance in Empirical Economics," *American Economic Review: Insights,* 2(2), 193–208.

American Evaluation Association. (2018). Guiding Principles for Evaluators. https://www.eval.org/Portals/0/Docs/AEA_289398-18_GuidingPrinciples_Brochure_2.pdf

Besharov, D. J. (2009). "Presidential address: From the Great Society to continuous improvement government: Shifting from “does it work?” to “what would make it better?”," *Journal of Policy Analysis and Management,* 28(2), 199-220.

Brodeur, A., Cook, N., & Heyes, A. (2020). "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review,* 110(11), 3634–3660.

Brown, A. B., & Klerman, J. A. (2012). "Independent Evaluation:Insights from Public Accounting," *Evaluation Review,* 36(3), 186-219.

Chopra, F., Haaland, I., Roth, C., & Stegmann, A. (2023). "The Null Result Penalty," *The Economic Journal,* 134(657), 193-219.

DellaVigna, S., & Linos, E. (2022). "RCTs to Scale: Comprehensive Evidence From Two Nudge Units," *Econometrica,* 90(1), 81-116.

Fels, K. M. (2022). "Who nudges whom? Expert opinions on behavioural field experiments with public partners," *Behavioural Public Policy,* 9(1), 212-248.

Ioannidis, J. P. (2016). "Evidence-based medicine has been hijacked: a report to David Sackett," *Journal of Clinical Epidemiology,* 73, 82-86.

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). "Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention," *Trends in Cognitive Sciences,* 18(5), 235-241.

Jefferson, T., Jones, M. A., Doshi, P., Del Mar, C. B., Hama, R., Thompson, M. J., . . . et al. (2014). "Neuraminidase inhibitors for preventing and treating influenza in adults and children," *Cochrane Database of Systematic Reviews*(4).

Kirsch, I. (2014). "The emperor's new drugs: medication and placebo in the treatment of depression," *Handb Exp Pharmacol,* 225, 291-303.

List, J. A. (2024). "Optimally generate policy-based evidence before scaling," *Nature,* 626(7999), 491-499.

Maier, M., Bartoš, F., Stanley, T. D., Shanks, D. R., Harris, A. J. L., & Wagenmakers, E.-J. (2022). "No evidence for nudging after adjusting for publication bias," *Proceedings of the National Academy of Sciences,* 119(31), e2200300119.

McGauran, N., Wieseler, B., Kreis, J., Schüler, Y.-B., Kölsch, H., & Kaiser, T. (2010). "Reporting bias in medical research - a narrative review," *Trials,* 11(1), 37.

Mertens, S., Herberz, M., Hahnel, U. J. J., & Brosch, T. (2022). "The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains," *Proceedings of the National Academy of Sciences,* 119(1), e2107346118.

Morris, M., & Clark, B. (2013). "You Want Me to Do What? Evaluators and the Pressure to Misrepresent Findings," *American Journal of Evaluation,* 34(1), 57-70.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). "The preregistration revolution," *Proceedings of the National Academy of Sciences,* 115(11), 2600-2606.

O'Hara, A. (2020). Model Data Use Agreements: A Practical Guide. In S. Cole, I. Dhaliwal, A. Sautmann, & L. Vilhuber (Eds.), *Handbook on Using Administrative Data for Research and Evidence-based Policy*.

Peck, L. R. (2018). "The Big Evaluation Enterprises in the United States," *New Directions for Evaluation,* 2018(160), 97-124.

Petrosino, A., Turpin-Petrosino, C., Hollis-Peel, M. E., & Lavenberg, J. G. (2013a). "Scared Straight and Other Juvenile Awareness Programs for

Preventing Juvenile Delinquency: A Systematic Review," *Campbell Systematic Reviews,* 9(1), 1-55.

Petrosino, A., Turpin - Petrosino, C., Hollis - Peel, M. E., & Lavenberg, J. G. (2013b). "'Scared Straight' and other juvenile awareness programs for preventing juvenile delinquency," *Cochrane Database of Systematic Reviews*(4).

Pieper, D., & Rombey, T. (2022). "Where to prospectively register a systematic review," *Systematic Reviews,* 11(1), 8.

Pleger, L., & Sager, F. (2018). "Betterment, undermining, support and distortion: A heuristic model for the analysis of pressure on evaluators," *Evaluation and Program Planning,* 69, 166-172.

Rossi, P. H. (2003). "*The iron law of evaluation reconsidered."* Paper presented at the Remarks presented at the 2003 APPAM research conference, Washington, DC.

Sawada, T., Yamada, H., Dahlöf, B., & Matsubara, H. (2009). "Effects of valsartan on morbidity and mortality in uncontrolled hypertensive patients with high cardiovascular risks: KYOTO HEART Study," *European Heart Journal,* 30(20), 2461-2469.

Schmidli, F. H., Pleger, L. E., & Hadorn, S. (2023). "Don't you forget about me: Independence of evaluations from the perspective of US evaluation clients—An exploratory study," *Evaluation,* 29(1), 110-132.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

Thomas, G. (2021). "Experiment's persistent failure in education inquiry, and why it keeps failing," *British Educational Research Journal,* 47(3), 501-519.

Tsang, B. (2025). "P hacking -Five ways it could happen to you," *Nature*.

U.S. Department of Health and Human Services. (2010). *Head Start Impact Study: Final report*. https://acf.gov/sites/default/files/documents/opre/hs_impact_study_final.pdf

US Office of Management Budget. (2020). "Phase 4 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Program Evaluation Standards and Practices," *Memorandum for the Heads of Executive Departments and Agencies*.

Weiss, C. H. (1993). "Where Politics and Evaluation Research Meet,"

Evaluation Practice, 14(1), 93-106.
伊芸研吾. (2024). エビデンス業界の市場分析. *経済セミナー, 2024 年 8・9 月号*.
関沢洋一. (2022). 医療における EBM から EBPM が学べること：ハイジャック問題を中心にした考察. *大竹文雄・内山融・小林庸平編著　EBPM：エビデンスに基づく政策形成の導入と実践*: 日本経済新聞出版.
植原亮. (2025). 科学的思考入門: 講談社.
西出順郎. (2020). *政策はなぜ検証できないのか：政策評価制度の研究*: 勁草書房.
西出順郎. (2023). "政策評価研究の黄昏," *ガバナンス研究,* 19, 19-33.
長谷川龍樹・多田奏恵・米満文哉・池田鮎美・山田祐樹・高橋康介・近藤洋史. (2021). "実証的研究の事前登録の現状と実践:OSF 事前登録チュートリアル," *心理学研究,* 92(3), 188-196.