RIETI Policy Discussion Paper Series 26-P-011

EBPM（エビデンスに基づく政策形成）入門
第4話　設計は分析に勝る（効果検証についての補足情報）

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-011
2026 年 5 月

EBPM（エビデンスに基づく政策形成）入門[1]

第４話　設計は分析に勝る（効果検証についての補足情報）

関沢洋一（独立行政法人経済産業研究所）

要　旨

- 効果検証を理解する上で注意すべきこととして、①統計学上の概念で効果検証では頻繁に登場する「有意差」は万能ではない、②一度示された効果が再現されない場合がある、③効果があるかどうかが個体（人、企業等）によって異なる場合がある。
- 効果検証手法として操作変数法と傾向スコアマッチングがしばしば使われるが、分析結果を信用していいかどうかわからない場合があり、政策現場にいる人たちは慎重さを持ってこれらの手法に接することが望ましい。
- 効果検証を行うためには政策介入の設計段階からの準備が重要で、設計は分析に勝る。先行的に介入を受ける群とその間は待っている群を作る、単純に希望者全員を施策の対象とする制度設計は避けるなど、政策の設計に携わる人々が留意することが望まれるポイントがある。

キーワード：有意差、再現可能性、外的妥当性、効果の異質性、設計は分析に勝る
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４話　設計は分析に勝る（効果検証についての補足情報）

（要約）

- 効果検証を理解する上で注意すべきこととして、①統計学上の概念で効果検証では頻繁に登場する「有意差」は万能ではない、②一度示された効果が再現されない場合がある、③効果があるかどうかが個体（人、企業等）によって異なる場合がある。
- 効果検証手法として操作変数法と傾向スコアマッチングがしばしば使われるが、分析結果を信用していいかどうかわからない場合があり、政策現場にいる人たちは慎重さを持ってこれらの手法に接することが望ましい。
- 効果検証を行うためには政策介入の設計段階からの準備が重要で、設計は分析に勝る。先行的に介入を受ける群とその間は待っている群を作る、単純に希望者全員を施策の対象とする制度設計は避けるなど、政策の設計に携わる人々が留意することが望まれるポイントがある。

この第４話は第２話と第３話の続きになります。第１節において効果検証を理解する上での留意点をいくつか述べます。第２節では、第２話と第３話で取り上げなかった効果検証の方法として操作変数法と傾向スコアマッチングを取り上げますが、これらの手法による報告に惑わされないようにという注意喚起という面が強いです。

第３節において、第２話と第３話で述べた効果検証の手法を実際に利用することができるようにするために、行政官や政治家のような政策現場にいる人々が読むことを念頭において、後々の効果検証を可能とするような政策介入の制度設計のあり方について述べます。

第１節　効果検証を理解する上での留意点

以下では効果検証を理解する上で注意すべき点として、①有意差が万能ではないこと、②一度示された効果が再現されない場合があること、③効果があるかどうかが個体（人、企業等）によって異なる場合があることについて取り上げます。

有意差は万能ではない

効果検証の文献を見ると、「有意」「有意差」といった言葉が頻繁に出てきます。統計学では、介入群と対照群の間に差がないという仮説（帰無仮説）を作って、その仮説に照らすと 5%未満の可能性でしか起きそうにないめずらしい差があると、有意差があることとなって、帰無仮説が誤りという判断が下され

ます。

効果検証研究においては、有意差が出るようにすることが目指されることが多いです。たとえば、医薬品の審査においてはその薬が審査機関によって承認されるかどうかに有意差の有無が影響し、ひいては製薬会社の株価にまで影響を及ぼします。一般の効果検証の研究でも研究者は有意差を示そうとして頑張ろうとすることが多いです。ただ、有意差で一喜一憂するのは良いことではありません。

効果検証においてはサンプルサイズが大きいほど 95%信頼区間の幅が小さくなり、p 値が小さくなり、有意差が出やすくなります。表４－１の例では介入群と対照群がそれぞれ 1000 人の場合と 10000 人の場合で比べていて、リスク比が同じであるにも関わらず 1000 人の場合は p 値＝0.13、10000 人の場合には p 値は 0.001 未満となっています。また、介入群と対照群がどちらも 1000 人でもイベント数が倍になれば、リスク比が同じでも有意差がでます。表４－１では１列目は p 値が 0.13、イベント数が倍になった３列目では p 値は 0.03 で、後者では信頼区間の幅が狭くなり、１をまたがなくなっています。

表４―１　リスク比が同じ場合のサンプルサイズに応じた信頼区間と p 値の違い

各群の総数	イベント数		リスク比	95%信頼区間	p 値
	介入群	対照群			
1000 人	80 人	100 人	0.8	0.60–1.07	0.13
10000 人	800 人	1000 人	0.8	0.73–0.88	<0.001
1000 人	160 人	200 人	0.8	0.65–0.98	0.03

有意差がないことは効果があるとは言えないことを意味し、効果がないことを積極的に意味するものではありません。特に、サンプルサイズが少ない場合にはサンプルサイズを増やせば有意差が出る可能性もあるので、効果なしとして切り捨てるのは問題があり、慎重に見極める必要があります。

また、有意であっても、効果量（効果の大きさのこと）が小さければ、実質的な効果があるとは言えないことにも留意する必要があります。

たとえば、介入によって体重が 100 グラム減ったとか、血圧が 1mmHg 下がったといった結果では、たとえ統計的に有意でも実質的な効果があるとは言えないです。実践的な有意差（practical significance）、臨床的な有意差（clinically significance）という言葉がありますが、統計的な有意差（statistical significance）だけに執着してはいけないという意味でよく使われます。特にビッグデータを分析すると小さな効果量でも有意差はでやすくなる

ので気をつける必要があります。

有意差については統計学者がその利用に疑問を呈する声明を出しています[2]。著名な経済学者からも、統計的に有意でない場合の方が有意である場合よりも有意義な情報を提供する場合があること[3]、有意差の有無を示すことよりも点推定値と信頼区間を示す方が望ましいこと[4]が指摘されています。特に、政策介入の効果検証においては、単に有意であるかどうかを示すよりは、どの程度の効果が見込まれるか（点推定値、リスク比など）と、不確実な程度はどれくらいか（信頼区間）を示す方が、情報の受け手となる政策立案や評価に携わる人々にとって有意義なものとなります。

一度示された効果が再現されない場合がある

ある集団を対象として効果検証が行われて効果があると判断された場合、別の集団についてその結果を一般化できるかという問題があり、外的妥当性と呼ばれます。ここでいう集団とは、性別、年齢層、人種、国、地域など様々なものが含まれます。RCT のように因果関係の検証としては強力な分析手法であっても（内的妥当性と呼ばれます）、外的妥当性があることは保証されていません。

EBM（エビデンスに基づく医療）の場合、人体の構造には極端な違いがないので、性別や年齢のような一部の例外を除くと外的妥当性を心配する必要はあまりなさそうです。これに対して EBPM の場合には、法制度や社会の慣習や経済状況など国毎に様々な違いがあるため、ある国の結果がそのまま自国に妥当するかどうかを慎重に見極める必要があります。また、同じ国の中でも、地域毎の違いや、時期の違い（好況時と不況時など）で、同じような効果検証でも結果が変わってくるかもしれません。こうした事情により、EBM に比べて EBPM では外的妥当性により注意することが必要になります。

また、似たような集団であっても、初期の効果検証と同様の結果がその後の効果検証で示されない場合があります。同じ介入についての効果検証であっても、初期に行われるものはその後に行われるものの 2.67 倍高い効果（95% 信頼区間, 2.12-3.37）が示されるという研究があります[5]。また、ある政策介入について少人数を対象として効果検証を行った場合には効果があったものの、同じ集団を対象としてサンプルサイズを増やしていくと効果が小さくなる傾向が

---

[2] Amrhein et al. (2019)
[3] Abadie (2020)
[4] Imbens (2021); Romer (2020)
[5] Alahdab et al. (2018)

あることが指摘されています[6]。ボルテージドロップという言葉が使われます[7]。

これらのことはEBPMを進める上で課題になります。たとえば、どこかの市町村において政策介入に効果があったという報告がなされたとしても、それに飛びついていきなり全国レベルでその政策介入を行うことには慎重になる必要があります。

平均効果と異質性

介入の効果として示されるものは通常は平均値であり、その介入を受ける全ての人々に対して同様の効果が生じることを意味するわけではありません。たとえば、第2話で紹介した不眠に対する認知行動療法等の介入（第2話の図2－2）では全ての人々に効果があったわけではなく、アウトカム変数である不眠尺度（PSQI）が介入後に改善しなかった人々もいます。効果検証の結果を説明する文章ではいかにも全員に効果がありそうな印象を与えることがありますが、全員に効果があることはまずないです。

介入の対象者のうち一部の集団には効果があるものの他の集団には効果がありそうにない場合や、効果の大きさが違ったりする場合があり、効果の異質性と呼ばれます。第3話で取り上げた企業向けの研究開発補助金のRDD（第3話の表3－1）は効果の異質性の例であり、大企業と中小企業で効果の方向性が反対となり、全体の効果が相殺されていました。

近年の研究では機械学習を使って効果の異質性を検証する動きが出ています[8]。たとえば、平均的に見ると効果が観察できなかったけれども、この集団（例えば、若年層、女性）に対しては効果があったという分析結果を示せるようになってきました。

どの集団に介入の効果があったかわかれば、その集団にターゲットを絞ることによってリソースの無駄遣いが避けられることになり、異質性を尊重する動きは望ましいと思います。

第2節　慎重さを持って接することが望ましい効果検証の手法

EBPMに興味を持つ人がしばしば目にする分析手法として、第2話と第3話で述べたRCT、RDD、DIDの他、操作変数法や傾向スコアマッチングがあります。個人的にはこれらの手法を行政の現場がどの程度信頼していいかについ

---

[6] Indig et al. (2017); Ioannidis et al. (2021)
[7] McKay et al. (2023); リスト(2023)
[8] Wager and Athey (2018)

ての疑問があり、政策現場にいる人たちが慎重さを持って接することが望ましい効果検証手法として簡単に説明します[9]。

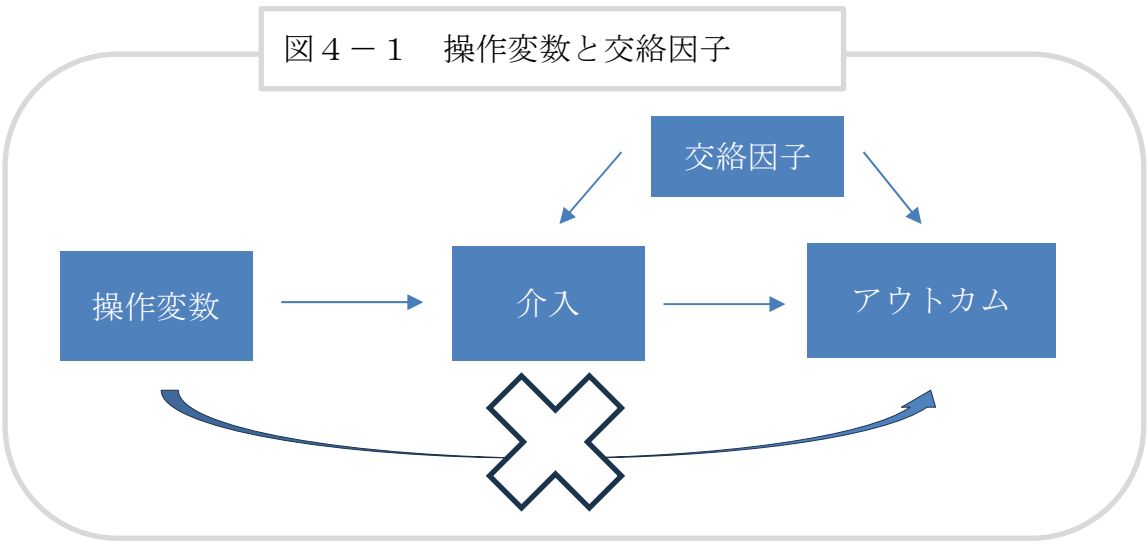
操作変数法

操作変数は介入には影響を及ぼすものの、アウトカムには介入を通じてしか影響を及ぼさない変数を指し、操作変数があると因果関係の検証が可能になることがあります。イメージすると図４－１になります。

典型的な操作変数はRCTによる割り付けの変数です。この場合、コイントス（またはそれに類似するもの）が生み出した０と１の数値（これが操作変数）は、この数値が１になった人々は介入群（本当の介入を受ける）になり、０になると対照群（プラシボ介入を受ける）になることを通じて、アウトカムに影響を及ぼします。

特に、RCTの中でも、どちらのグループになったかが研究の現場にいる医療関係者も患者もわからない二重盲検法においては、介入効果以外の経路でアウトカムに及ぼす影響が遮断されます（図４－１の☒。）

図４－１　操作変数と交絡因子



操作変数法を使った研究は多くありますが、使用した操作変数が適切であること（特に介入以外の経路で操作変数がアウトカムに影響を及ぼさないこと）をデータから直接検証することは困難です[10]。このため、分析結果にアクセス

[9] ここでの私の指摘と似たことは既に伊藤・金子（2024）で述べられています。

[10] RCT（第２話で取り上げた奨励デザインを含みます）、くじびきによる割当、第３話で取り上げたファジーRDD（操作変数法として扱われます）は操作変数法が適切となる例です。ただし、その適切さは割当の偶然性など研究の設

する側としては、操作変数法を使用した分析の多くについてその操作変数が本当に妥当なのかの確信を持てず、モヤモヤ感が残ることが多いです。

傾向スコアマッチング(Propensity score matching)

RCT のような実験を伴わずにデータだけが集まった研究（観察研究）では、介入とアウトカムの双方に影響を及ぼす変数（計測されていないものも含む）が存在するのが普通で、交絡因子と呼ばれます（図４－１、第１話の図１－２、第２話の図２－１）。重要な交絡因子は全て調整しないと観察研究において因果関係の特定はできません。

重要な交絡因子となる変数を用いて因果関係を特定しようとする主な取り組みとして、傾向スコアマッチングがあります。

交絡因子を中心に観測された変数を使って、それぞれの者が介入を受ける確率（傾向スコア）を計算します。介入群の個々の者と傾向スコアがほぼ同じ対照群の者をペアにする（マッチさせる）と、傾向スコアを計算するために使った変数において、介入群と対照群の属性が揃ってきて、あたかも「疑似ランダム化」したように介入群と対照群の比較が行えるようになります。似たような手法として逆確率重み付け（Inverse Probability Weighting）という手法があり、こちらも傾向スコアを利用して介入群と対照群の属性を揃えることを目指します。

「疑似ランダム化」が成功するためには交絡因子のうち重要なものがすべて計測されていてマッチングに使えることが必要です。残念ながら、実際には交絡因子となる変数の全ては計測されていない場合が多いです（やる気や真面目さのような心理的な変数など）。この場合、傾向スコアに基づく分析そのものは行えても分析結果が真の値（通常はわからない）からかけ離れる（バイアスが生じる）ことが懸念されます。

実際には既にあるデータを入手してそれだけを使って分析する場合が多くて、このような場合には、利用できる変数も限られているため、分析そのものをあきらめるか、存在する変数の範囲で傾向スコアを算出せざるを得なくなります。データの収集や整理に費やした努力を無駄にしないためや、とにかく結果を出す必要性に駆られて、（本当は行うべきではないにも関わらず）傾向スコアに基づく分析を強行することも多いように思います。こういう事情があるため、分析結果が本当の効果からかけ離れる（バイアスが生じる）可能性は拭い去れず、操作変数法と同じようにモヤモヤ感が残る場合が多いです。

計に基づくもので、観測データから決まってくるものではありません。

第３節　後々の効果検証に向けた政策介入の制度設計のための参考情報

ここからは後々で効果検証できるようにするための政策介入の設計段階からの準備について述べます。因果推論の創始者であるドナルド・ルービンが「客観的な因果推論のためには、設計が分析に勝る（For objective causal inference, design trumps analysis.）」と述べており、事前準備の重要性を強調しています[11]。

比較可能な介入群と対照群を作る

効果検証を行うためには、ある介入を受けた介入群（人、企業、地方公共団体など）と受けなかった対照群が存在することが必要になります。A という介入を受けたグループと B という介入を受けたグループという場合も、２つの介入の間の比較が可能です。これに対して、全ての人々（企業、地方公共団体など）が対象となる介入を一斉に行うと、比較対象が存在しないため、効果検証は難しくなります[12]。

もう１つ重要なことがあります。介入群と対照群は似たような集団であることが望ましいです[13]。この言葉を文字通り捉えると効果検証はとても難しく思えますが、そんなことはないです。以下で説明していきます。

①数値に基づいて政策介入の対象を決めると効果検証を行いやすい

個人や法人がある施策の対象となるか否かを決めるに当たっては、あらかじめ定められた数値を超える場合のみを対象としたり、数値を超えた場合のみに二次審査の対象としたりする場合が多いです。このような取り扱いは後に介入の効果検証を行う上で役に立ちます。というのは、基準となる数値の周辺においては介入の対象となる者は介入の対象とならない者と似たような属性を持つようになり、RCT に似た状況になるためです。これにより RDD の適用が可能になります。

たとえば、企業を対象とする補助金の多くは申請企業の申請書類を踏まえて事務局が得点を作って、その得点に従って採択を決めています。メタボ健診の保健指導の対象者は腹囲 85cm 以上（女性は 90cm 以上）などを基準として決められます。ストレスチェックを行う義務のある事業所は、従業員数 50 名以

---

11 Rubin (2008)

12 前後比較で効果検証できる場合もありますが(Kontopantelis et al., 2015)、エビデンスとしては弱くなります。

13 DID では介入群と対照群が似たような集団であることは求められませんが、アウトカムとなる変数が介入群の介入以前に両群において概ね平行に推移することが求められます（第３話参照）。

上と定められています（但し 2025 年の法改正で全ての事業所に対象を広げることが決まりました)。学校の入試の多くは合格者を一定の点数以上としています。施策の対象となる希望者に順位を付けて、上位 3 割とか、上位 1000 名とか、順位に応じて施策の対象を決める場合もあります。

気を付けるべきことがあります。ある様々な介入を行う基準となる数値がいろいろな介入で同一になる場合は、複数の介入のどれによる効果がわからなくなることがあります。たとえば、従業員数 50 名以上では産業医の選任、ストレスチェックなど、様々な施策がこの数値を使っていて、どの介入の効果を検証しているのかがわからなくなります。事後的に効果検証を行えるようにするためには、介入の基準となる数字を慎重に決める必要があります。

②抽選が必要な場合、ランダム化によるくじびきを行う

ある施策の対象となる人数が限られている場合、抽選で対象者を選ぶことがあります。この場合、操作変数法による分析が可能になりますし、RCT に持ち込むこともできます。

③積極的に募集をかける場合には奨励デザインに持ち込む

募集者数に対して参加者数が足りなさそうで、何らかの宣伝が必要になる場合には、たとえば電話で勧誘する参加者リストを作成して、そのリストの作成の際に偶然の要素を盛り込むと効果検証が行いやすくなります（電話番号の末尾が偶数の企業にのみ電話するなど)。生真面目さを捨てて何らかの偶然的要素を盛り込むことが、効果のある政策介入かどうかを見極めるという価値を生むことになります。

たとえば、メタボ健診に基づく保健指導の対象者に対して、通常の連絡に加えて電話などで保健指導への参加を促す場合に、対象者全員に行うのではなく、ランダム化して選んだ人々にだけ参加を促すことにします（第 2 話で取り上げた奨励デザインになります)。ランダム化だと問題があると感じられる場合には、健診結果のデータを使って循環器疾患リスクを数値化した指標[14]を作ってその数値が一定値を超える人々に対してのみ電話することとすれば RDD の適用が可能になります。

④先行的に介入を受ける群とその間は待っている群を作る

政策介入の効果が比較的早く起きることが予想される場合、先行的に介入を

[14] 一例として、イギリスで広く用いられている循環器疾患リスク指標である QRISK があります(Hippisley-Cox et al., 2017)。

受ける群とその間は待っている群を作ることが考えられます。

たとえば、残業減少を目的とする課長研修が課員の残業を減らすかどうかを検証したいとします。この場合、課長を２つのグループにランダムに分けて、先行研修群と待機群に分けて、先行研修群はすぐに研修を受けてもらって、待機群は半年後に研修を受けてもらうようにします。そうすると、待機群が研修を受けるまでの半年間の間に研修を受けた課長のいる課の残業時間が研修によって減ったかどうかの検証が行えます。

待機群を作る場合は、待機群をあまり待たせるわけにはいかないので、効果が数か月ぐらいで現れるものを中心に考える必要があるという制約はあります。

⑤単純に希望者全員を施策の対象とする制度設計は避ける

ある施策について事後的な効果検証を行う必要がある場合は、効果があることが明らかになるまでは、希望者全員を施策の対象としない方がいいです。自己選択バイアスという問題が生じます。

この点で問題がある例としてメタボ健診があります（第５話参照）。メタボ健診の実施率は概ね６割ですので、比較可能な介入群と対照群が存在するように見えます。しかし、メタボ健診を受けた人々は自分で受けると決めた人々です。この場合に自己選択バイアスが生じます。メタボ健診を受ける人々はもともと健康志向が高いなど、受けない人々と最初から違いがあり、信頼できる分析の妨げになります。

この場合、上記の①から④までのいずれかの持ち込むことが望まれます。

サンプルサイズを減らさないように気を付ける

分析を行うに当たってはサンプルサイズが多い方が結果のぶれが少なくなり、信頼できる結果を得やすいです。統計学的な有意差はサンプルサイズが多い方が出やすくなるので、有意差で効果を示したいのであれば、サンプルサイズが多い方が望ましいです。

また、全体としてのサンプルサイズが同じであっても、それが分割されていると実際のサンプルサイズが少なくなり、分析が難しくなります。例えば、中小企業向けの補助金ではグリーン枠、デジタル枠などいろいろな枠が作られていて、それぞれのサンプルサイズが少なくなるので有意差がなかったと報告せざるを得ない分析結果が出やすくなります。

市区町村で行う効果検証の場合には、政令指定都市のような多くの人口がある場合を除くと、どうしてもサンプルサイズが小さくなります。複数の市区町村が連携して効果検証に取り組めるようになることが望まれます。

同じ施策が繰り返し行われることを避ける

同じ施策が繰り返し行われると効果検証が行いにくくなります。たとえば、中小企業の補助金は年に数回の公募があり、しかも年ごとに公募が行われます。この場合、最初の公募で不採択となった企業が次回以降の採択で採択されることがしばしばあります。そうすると、仮に最初の公募の効果検証をしようとすると、採択された企業と不採択の企業の比較をしようとしても、実際には最初に採択された企業と後で採択された企業の比較に近くなり、正確な効果検証の妨げになります。

別の例として、メタボ健診の保健指導について、ある年に保健指導を受けた場合の効果検証をしようとすると、翌年の効果検証は行えるのですが、２年より先になると、最初の年に保健指導を受けなかった人の中に翌年に保健指導を受けた人が出てくるため、２年後以降の効果検証は難しくなります（第５話参照)。

引用文献

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

Alahdab, F., Farah, W., Almasri, J., Barrionuevo, P., Zaiem, F., Benkhadra, R., . . . Wang, Z. (2018). "Treatment Effect in Earlier Trials of Patients With Chronic Medical Conditions: A Meta-Epidemiologic Study," *Mayo Clinic Proceedings,* 93(3), 278-283.

Amrhein, V., Greenland, S., & McShane, B. (2019). "Scientists rise up against statistical significance," *Nature,* 567(7748), 305-307.

Hippisley-Cox, J., Coupland, C., & Brindle, P. (2017). "Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study," *BMJ,* 357, j2099.

Imbens, G. W. (2021). "Statistical Significance, p-Values, and the Reporting of Uncertainty," *The Journal of Economic Perspectives,* 35(3), 157-174.

Indig, D., Lee, K., Grunseit, A., Milat, A., & Bauman, A. (2017). "Pathways for scaling up public health interventions," *BMC Public Health,* 18(1), 68.

Ioannidis, J. P., Maniadis, Z., & Tufano, F. (2021). When is evidence actionable? Assessing whether a program is ready to scale *The Scale-Up Effect in Early Childhood and Public Policy* (pp. 126-142):

Routledge.

Kontopantelis, E., Doran, T., Springate, D. A., Buchan, I., & Reeves, D. (2015). "Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis," *BMJ,* 350, h2750.

McKay, H. A., Macdonald, H. M., Nettlefold, L., Weatherson, K., Gray, S. M., Bauman, A., . . . Sims Gould, J. (2023). "What is the 'voltage drop' when an effective health promoting intervention for older adults-Choose to Move (Phase 3)-Is implemented at broad scale?," *PLOS ONE,* 18(5), e0268164.

Romer, D. (2020). "In Praise of Confidence Intervals," *AEA Papers and Proceedings,* 110, 55–60.

Rubin, D. B. (2008). "For Objective Causal Inference, Design Trumps Analysis," *The Annals of Applied Statistics,* 2(3), 808-840.

Wager, S., & Athey, S. (2018). "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association,* 113(523), 1228-1242.

リスト，ジョン／高遠裕子訳. (2023). *『そのビジネス、経済学でスケールできます』*：東洋経済新報社.

伊藤寛武・金子雄祐. (2024). Python で学ぶ効果検証入門: オーム社.