RIETI Policy Discussion Paper Series 26-P-005

EBPM（エビデンスに基づく政策形成）入門
第2話　効果検証の重要な要素とランダム化比較試験

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-005
2026 年 3 月

EBPM(エビデンスに基づく政策形成)入門[1]

第２話　効果検証の重要な要素とランダム化比較試験

関沢洋一（独立行政法人経済産業研究所）

要　旨

・信頼できるエビデンスを構築するためには、①比較可能な介入群と対照群の存在、②計測可能なアウトカムの存在の 2 つが重要になる。
・ランダム化比較試験（RCT）では、シンプルなコイントスによって介入群と対照群を分けることによって、比較可能な介入群と対照群を作ることができる。
・RCT の応用である奨励デザインでは政策介入の対象となりうる人々をランダムに奨励群と対照群に分けた上で、奨励群のみに政策介入への参加を促す。
・クラスターRCT では、ランダム化の単位を教室や学校や地域などの集団とすることによって、個人単位での RCT が難しい場合に対処することが可能になる。
・有意差や 95%信頼区間など統計学に出てくる言葉を知っていると効果検証の分析結果を理解しやすくなる。

キーワード：介入群、対照群、アウトカム、有意差、９５％信頼区間、奨励デザイン、クラスターRCT

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 2 話　効果検証の重要な要素とランダム化比較試験

（要約）

- 信頼できるエビデンスを構築するためには、①比較可能な介入群と対照群の存在、②計測可能なアウトカムの存在の 2 つが重要になる。
- ランダム化比較試験（RCT）では、シンプルなコイントスによって介入群と対照群を分けることによって、比較可能な介入群と対照群を作ることができる。
- RCT の応用である奨励デザインでは政策介入の対象となりうる人々をランダムに奨励群と対照群に分けた上で、奨励群のみに政策介入への参加を促す。
- クラスターRCT では、ランダム化の単位を教室や学校や地域などの集団とすることによって、個人単位での RCT が難しい場合に対処することが可能になる。
- 有意差や 95%信頼区間など統計学に出てくる言葉を知っていると効果検証の分析結果を理解しやすくなる。

本稿では第１節で効果検証において重要な２つの要素について述べます。第２節において効果検証の黄金律とも呼ばれるランダム化比較試験（RCT）について取り上げます。

本稿（第２話）から第４話までで目指しているのは、自分の担当する政策介入の効果を検証するためにこの手法を使えるのではないか、効果検証を後々で行うためには自分の担当する政策介入をこういう設計にした方がいいのではないかというイメージを政策の立案や実施に携わる現場にいる人々につかんでもらうことです。また、研究者や民間シンクタンクが行う効果検証の分析結果を、専門的知識を持たないものの政策介入の効果検証に関心を持つ人々が自ら批判的に吟味するための基礎的な情報を提供することも目指しています。

第１節　効果検証で重要な２つの要素

信頼できるエビデンスを構築する上で重要になるのは、①比較可能な介入群と対照群の存在、そして、②計測可能なアウトカムの存在の 2 つです。

比較可能な介入群と対照群の存在

政策介入の効果検証によって構築されるエビデンスが信頼できるためには、何らかの介入を受ける個体（人、企業、地方公共団体など）の集まり（介入群）とその介入を受けない個体の集まり（対照群）があり、かつ、その 2 つの

集まりが比較可能になっていることが求められます。この点が厳密に守られるほどエビデンスとしての信頼度が上がります。

このことは表２－１に分析手法のエビデンスのレベルとして示されています。一番上のレベル５になっているのが次節で説明する RCT であり、RCT は介入を受けるグループと受けないグループが存在するだけでなく、後述するとおり２つのグループの属性が似たものになっていて容易に比較可能になっています。

次にレベル４として回帰不連続デザイン（RDD）と抽選を使った操作変数法が挙げられています。いずれも RCT ほどではないものの、比較可能な 2 つのグループが存在します。RDD は第３話で、操作変数法は第４話で触れます。

次にレベル３として差の差分析（DID）が挙げられています。DID は２つのグループが似ているとは言いがたいのですが、平行トレンドという仮定を満たすという前提の下で、２つのグループが比較可能であると扱われます。詳しくは第３話で紹介します。

表２－１で何とかレベル３以上に持ち込むのが政策介入の効果検証に当たっての基本方針となります。

レベル２以下は、効果検証の実例として頻繁に見るものの、信頼度が低い効果検証です。政策介入の対象となった者に特化して行われる事後調査（セミナー参加者や補助金受給者への満足度調査など）や介入前後の比較は効果検証としては信頼度の低いものになります。

表２－１　エビデンスのレベル

レベルの高低	利用データと分析手法の概要
5(最も高い)	ランダム化比較試験（RCT）
4	実験に類する状況を利用した分析（回帰不連続デザイン[RDD]、抽選を使った操作変数法など）
3	実験はなく（観察のみ）、介入群と対照群の両方の介入前後のデータを用いた分析（差の差分析[DID]など）
2	① 介入群だけの前後比較（コントロール変数あり） ② 介入群と対照群の１時点（介入後）のデータだけを利用した分析（コントロール変数あり）
1(最も低い)	① 介入群だけの前後比較（コントロール変数なし） ② 介入群と対照群の１時点（介入後）のデータだけを利用した分析（コントロール変数なし）

（出典）What Works Centre for Local Economic Growth (2016)を元に作成。

計測可能なアウトカムの存在

政策介入の効果検証を行う上でもう一つ重要なのはアウトカムです。何らかの介入がある場合、その目的の達成度を明らかにする何らかの数値が存在しており、これはアウトカムと呼ばれます。アウトカムとは例えば次のものです。ある市が市民の血圧を下げることを目的として各家庭に血圧計を配布するプログラムを行った場合、典型的なアウトカムは市民の血圧で、これは様々な数値をとる連続変数です。これに対して、市民の血圧を計測した結果、高血圧（上の血圧が 130 以上または下の血圧が 85 以上）と判断される場合を 1、高血圧でないと判断される場合を 0 とする変数（二値変数と呼ばれます）をアウトカムとして設定することができます。二値変数を使うことによってデータとは思えないような様々な出来事をアウトカムとして数値化できます。

効果検証を行うためにはアウトカムが実際に計測できるものであることが必要です。アウトカムの実際の数値を取得するために、研究実施者が追跡調査を行って介入群と対照群から回答してもらう場合もありますが、時間が経過するにつれて回答してもらえなくなる場合が増えるという限界があります。また、補助金のような便益を与える措置では介入群では追跡調査を行えても、補助金をもらわない対照群は望み薄です。大部分は回答してもらえないです。

EBPM は政策介入の効果を検証するものなので、政策の実施主体である政府の保有する業務情報を使うことを検討することが望ましいですし、これは世界的な潮流でもあります。効果検証を行うに当たってはアウトカムとすることのできるデータが役所の取得した情報の中にあるかどうか、ある場合にはそれを利用することができるかどうかを確認することが必要になります。ここで述べたデータの話は極めて重要ですので、第 6 話で詳細に触れることとします。

第２節　ランダム化比較試験（RCT, randomized controlled trial）[2]

ランダム化比較試験（RCT）は介入とアウトカムの間の因果関係を検証する上で最も強力で信頼できる実験手法です。RCT は医薬品の市場投入の認可を受けるための試験で使われており、現代医療に大きく貢献しています。また、社会科学における RCT の活用の拡大が進んでいます。このような強力な手法でありながら RCT は小学校の理科の実験にも似たシンプルさを有しています。

一例として、A さんがビタミン D のサプリを継続的に摂取することによって骨折のリスクを減らせるかどうかを知りたいとします。この問いに答えてビタ

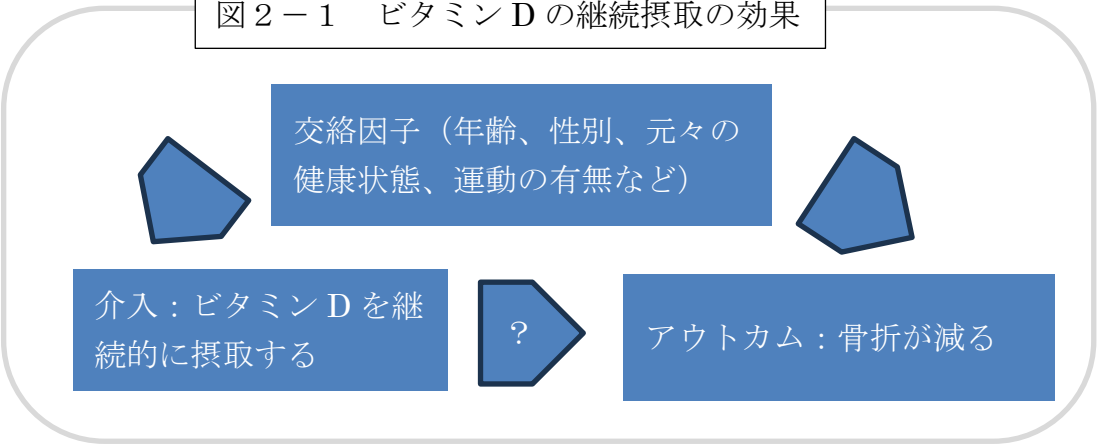
[2] RCT の重要性を一般向けに明らかにした解説書としてリー(2020)は有意義です。

ミン D サプリの摂取と骨折リスクの間の因果関係を正確に把握するためには、A さんがこのサプリを摂取しない場合と比較する必要があります。

しかし、現実世界では同一人物において両方の状況を同時に観察することは不可能です。タイムマシンを使わない限り、A さんがこのサプリを継続的に摂取する場合と摂取しなかった場合を直接比較することはできません。

次善の策として考えられることはビタミン D のサプリを摂取する人々と摂取しない人々を比較することです。ところが、単純に比較すると、ビタミン D の摂取と骨折の両方に影響を及ぼす要因[3]（年齢、性別、元々の健康状態、運動の有無などの交絡因子）が正確な効果検証の妨げになります（図２－１）。交絡因子とは介入とアウトカムの双方に影響を及ぼす要因を指します。

図２－１　ビタミン D の継続摂取の効果



介入の因果効果を推定するためには、A さんと似たような人々の集まりに着目します（グループ A）。たとえば、グループ A に属する人々が中高年の男女だとします。グループ A に属する人々を２万人集めてきて、一人一人についてコイントス[4]で表が出た場合には継続的にビタミン D を摂取してもらい（グループ A1）、裏が出た場合には同じ期間中に摂取しないでもらうこと（グループ A2）が可能だったとします。これが典型的なランダム化比較試験（RCT）です。

このコイントスによって、不思議なことにグループ A１とグループ A2 のどちらも１万人に近い数字になります。更に、特別な調整を行わなくても、年齢、男女比、元々の健康状態、運動の有無など様々な指標の分布が２つのグループで似たようなものになります。つまり、グループ A１と A2 はコイントス

[3] McCormack et al. (2017)。
[4] 実際にはパソコン上でコイントスに似たことを行えます。

という偶然の力を借りるだけで似たような属性を持つ集団となり、比較が可能になります。

RCT の実例１（アウトカムが二値変数）

ビタミン D サプリの継続摂取が骨折の予防につながるかどうかを検証するために行われた大規模な RCT（以下では VITAL）があります[5]。

VITAL では 25871 名の中高年の男女がランダムに２つの集団に割り当てられ、ビタミン D サプリの継続摂取に割り当てられたグループ（介入群）が 12927 名、プラセボ（ビタミン D の成分を有しない偽サプリ）を継続摂取するグループ（対照群）が 12944 名とほぼ同数になっています。

約 5 年間の追跡期間内に骨折を経験した人数は介入群では 769 名で、対照群では 782 名でした（表２－２）。

表２－２　ビタミン D サプリの継続摂取のランダム化比較試験（VITAL）

	サプリを摂取（介入群）	プラセボ摂取（対照群）
骨折あり	769 (a)	782 (b)
研究参加者数	12927 (c)	12944 (d)

（出典）LeBoff et al. (2022)

VITAL の場合、アウトカムとなる骨折は、発生した場合には１、発生しない場合には０となる二値変数になります。二値変数の場合、アウトカムとなる出来事（この場合は骨折が起きること）をイベントと呼ぶことがあります。

下記の式にあるとおり、介入群のイベント発生率（ここでは骨折）は 5.95% で対照群は 6.04%となっています。介入群のイベント発生率の方がわずかに低いのですが、これをもって効果ありとしていいのでしょうか。

$$リスク比(RR) = \frac{a/c}{b/d} = \frac{769/12927}{782/12944} = \frac{5.95\%}{6.04\%} \fallingdotseq 0.98$$

介入群のイベント発生率を対照群のイベント発生率で割った数値はリスク比と呼ばれます。この場合、上記の式にあるとおり、リスク比は約 0.98 になります。1−0.98 ＝ 0.02 が介入（この場合はビタミン D サプリの継続摂取）に

[5] LeBoff et al. (2022)。VITAL とは Vitamin D and Omega-3 Trial の略です。

よるリスクの減少割合で、この場合はビタミン D サプリを継続摂取した場合に骨折するリスクが相対的に 0.02（＝2%）低いことになります。

しかし、個々の RCT で観察されたリスク比は偶然に左右されます。VITAL と同様に中高年の男女（これは母集団と呼ばれます）の中から改めて VITAL と同じ人数を集めてきて仮に VITAL と同じ設定で同じような RCT を行っても、全く同じリスク比になることはほとんどあり得ず、同じような RCT を何度も行えば、毎回少しずつ違うリスク比が得られるのが普通です。そうすると VITAL においてリスクが 2%減ったというだけで効果があると判断するのは望ましくないことになります。

この点に対応するために、統計学では 95%信頼区間という範囲が計算されます。表２－２から計算した結果では 95%信頼区間の下限が 0.89、上限が 1.08 となり、リスク比（95% 信頼区間）が 0.98（0.89～1.08）と表記されます。

RCT のように厳密な効果検証手法が用いられるという前提の下ですが、よく見られる慣行としては、95%信頼区間の上限が 1 を下回れば（例えば、0.89～0.99）、有意差があって効果あり、95%信頼区間の下限が 1 を超えれば（例えば、1.01～1.11）、有意差があって負の効果ありと判断されます。95%信頼区間が 1 をまたぐときには、有意差はなく効果があるとは言えないと判断されます。VITAL の場合、95%信頼区間は 0.89～1.08 で 1 をまたぐので、有意差はなくて効果があるとは言えないということになります。

突然に「有意」という言葉がでてきましたが、これは日常的に使われる「重要度」や「意義」といった意味を示す言葉ではなく、統計学の専門用語で、統計的に見た「めずらしさ」を指します[6]。専門家からは違うと言われるかもしれませんが、専門家でない人々が何となく理解する上では、有意差は偶然では説明できない差という捉え方で問題ないと思います。

有意差があるかないかを示すものとして、p 値という言葉もしばしば使われます。p 値が小さいほど偶然では説明できない違いがあると判断されやすくなり、特に p 値が 0.05 を下回ると両群には有意差があると判断されます。VITAL のリスク比について p 値は 0.75 と計算され、有意差があるとは言えません。

リスク比と似たような言葉として、オッズ比とハザード比があります。算出の仕方は異なりますが、いずれも解釈の仕方は似ていて、数値が 1 を下回り、かつ、95%信頼区間の上限が 1 を下回った場合（95%信頼区間が 1 をまたがない場合）に効果があると判断されます。VITAL の原論文ではハザード比で分析結果を報告しており、ハザード比が 0.98、95%信頼区間は 0.89～1.08 で、p

[6] 加納・浅子（1998）

値が 0.70 になっており、リスク比と似たような数値になっています。

RCT の実例２（アウトカムが連続変数）

VITAL の場合、アウトカム変数は二値変数でした。アウトカム変数が連続変数の場合には少し扱いが異なります。私も参加した研究を紹介します[7]。

薬を使うことなく心の健康を増進するための手法として認知行動療法（CBT, Cognitive Behavioral Therapy）という治療法があります。CBT はセラピストに頼ることなく、自分で本を読んで取り組んだりインターネットのプログラムで取り組んだりすることも可能になっています。また、同じく心の健康を増進するための取り組みとしてポジティブ心理学という学術分野が発達していて、その代表的な手法として、毎晩寝る前にその日に自分に起きた良いことを３つ書くというエクササイズ（TGT, Three Good Things）があります[8]。

無料のインターネットのプログラムで CBT や TGT を自分で行うことによって心の健康を増進することができれば、多くの人々のウェルビーイングの向上につながるだけでなく医療費の伸びの抑制にもつながる可能性があります。しかし、その前提としてこのようなプログラムに効果があることが必要です。

千葉大学と独立行政法人経済産業研究所の共同研究として、インターネット上で行われる CBT（ICBT）と TGT のエクササイズが不眠を改善する効果があるかどうかを検証する RCT が行われました[9]。

この研究は千葉大学からインターネット調査会社への委託契約に基づいて行われました。調査会社には多くのモニターが所属していて、委託元の依頼に基づいてモニターを対象としてオンラインによるアンケート調査を行っています。RCT では多数の参加者を集めることやアウトカムの計測が課題ですが、この課題に向き合うため、この研究では、調査会社のモニターの中から研究参加者を募ってアンケート調査形式でアウトカム（不眠の程度を表す PSQI [ピッツバーグ睡眠質問票]という数値）の計測を行いました。PSQI の点数の範囲は 0 点から 21 点で、数値が高いほど睡眠に問題があるとされます。

この研究では不眠の問題を抱えた成人 312 名をランダムに ICBT 群と TGT 群と待機群（介入を受けないグループ）の３つのグループに分けました。4 週間の介入期間を経て、介入期間直後の 4 週間後、フォローアップとしての 8 週間後に、PSQI の各群の差を分析しました。

分析の結果は図２－２に示しています。この場合、介入終了直後における

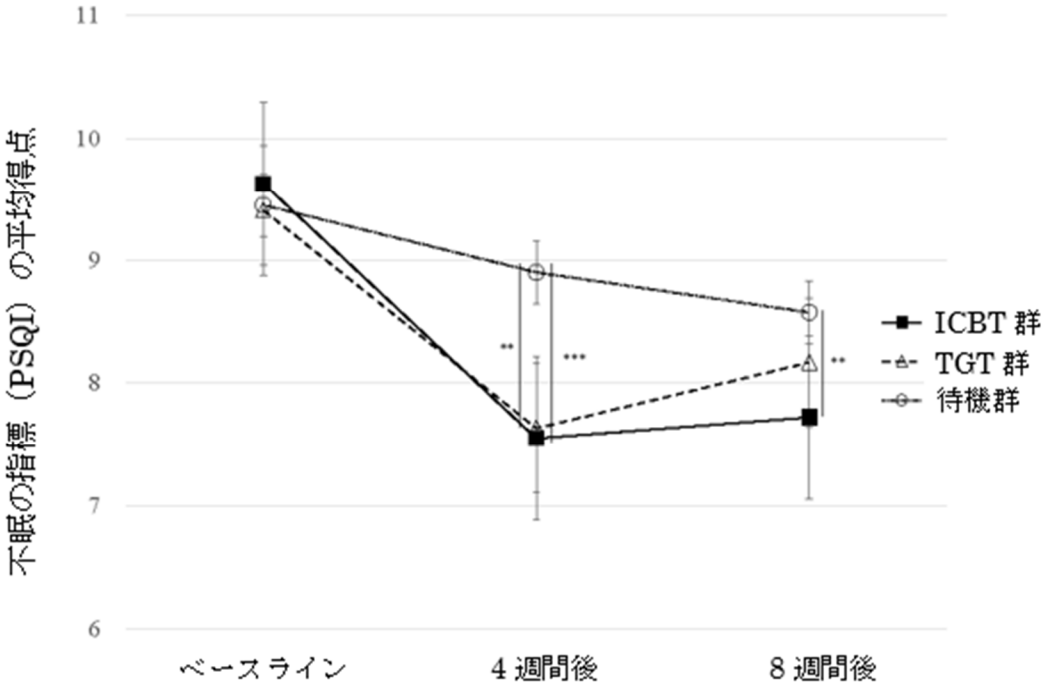
---

[7] Sato et al. (2022)
[8] Seligman et al. (2005)
[9] Sato et al. (2022)

ICBT 群と待機群の PSQI の平均差の点推定値（直接計測できない母集団の平均差として推定される値を１つの値で示したもの）は－1.56（ICBT 群の方が低い）で、95%信頼区間は－2.52～－0.59 になりました。連続変数の場合、95%信頼区間が 0 をまたがない場合に有意差があると判断されます。この研究の場合、95%信頼区間は 0 をまたいでおらず、数字がマイナスの場合に望ましい方向に向かったことになるので、効果があったと判断されることになります。有意差においても p＜0.001 となっていて（p は p 値のこと）、5%水準で効果があったと言えることになります。似たことを TGT 群について記述すると、TGT 群対待機群が、平均差の推定値が－1.15（95%信頼区間：－2.08～－0.23，p=0.0018）となり、こちらも効果があったと判断されることになります。

図２－２にはカタカナのエのような縦線があり、これはエラーバーと呼ばれます。ここでは 95%信頼区間が示されています。

図２－２　認知行動療法等が不眠の程度を示す指標（PSQI）に及ぼす影響



（出典）Sato et al. (2022)

ベースライン（介入前の数値）では 3 つの群のエラーバーが重なっていて 3 群であまり違いがありません。介入直後の 4 週間後では ICBT 群と TGT 群ではほとんど平均値が重なっていますが、待機群とは違いがあり、95%信頼区間が重なっておらず、有意差があることを示します。その一方で、フォローアッ

プ時点である介入開始から 8 週間後では TGT 群は待機群との間で有意差がなく、効果が維持されなかったことを示唆します。ICBT 群ではエラーバーは重なっておらず、効果が維持されています[10]。なお、この図にはアスタリスク（*）が書いてあります。この場合、*は p 値が 0.05 より小さいことを示し、**と***はそれぞれ p 値が 0.01、0.001 より小さいことを示します。

図２－２からもう１つ学べることがあります。介入を受けていない待機群の PSQI の得点がベースラインに比べると 4 週間後、8 週間後のいずれにおいても有意に改善していることです。対照群のない単なる前後比較の場合、このような自然な改善を介入の効果と勘違いする場合があります。対照群のある RCT を行うことが重要なことを示す一例です。

RCT の応用編１（奨励デザイン）

RCT では何らかの介入への参加者を募った上で、本物の介入を受けるグループと偽物の介入を受けるグループ（あるいは何の介入も受けないか、一定期間は待機するグループ）に参加希望者をランダムに分けて介入の効果を検証します。ただ、この方法だと介入を受けたい人々が本人の意に反して介入を受けられない場合があるという倫理的な問題が生じる可能性があります。また、介入群に割り当てられた人々の多くが実際には介入を受けない場合の処理が難しくなることがあります。

以下ではこれらの問題に対処するための RCT として奨励デザインを紹介します。

奨励デザインでは政策介入の対象となりうる人々をランダムに奨励群と対照群に分けた上で、奨励群のみに政策介入への参加を促します。

一例として、低所得者へのエネルギー料金の一部を補填する補助金プログラム（CARE）のエネルギー使用量に対する効果を検証したアメリカの研究を取り上げます[11]。

CARE は申請ベースとなっていて、このプログラムの適用条件に該当する人々が自ら申請することによって、エネルギー料金の補填のメリットを受けられました。このことを利用して、このプログラムの適用条件に該当する人々を

---

[10] 信頼区間がエラーバーとなっている場合、２つのエラーバーが重なっていなければ両者には有意差がありますが、重なっている場合には有意差がないとは限りません。TGT 群と待機群の 8 週間後のエラーバーは重なっていますが、これだけでは両群に有意差があるかどうかはわかりません。本文中に記載されている分析結果によると有意差はありませんでした。

[11] Hahn and Metcalfe (2021)。正確には奨励群には多少のバリエーションがあります。

ランダムに 2 群に分けて 1 つのグループのみ申請を促す手紙が送られました（こちらが奨励群）。対照群の人々は研究に参加していることを知りません。

奨励デザインの場合、対照群の人々は研究に参加していることを知らない場合が多いため、アウトカムの計測をどうするかが特に課題になります。CARE 研究は SoCalGas というガス会社が参加した研究なので、同社が計測したアウトカムを使っていると思われます。

奨励デザインにおいては主に 2 つの分析方法があります。1 つめは実際に介入に参加したかどうかに関わりなくランダムに割り当てられた全ての人々を分析対象とします。ITT（intention-to-treat）と呼ばれます。もう一つ LATE (local average treatment effect、局所平均処置効果)と呼ばれる推計があり、こちらは奨励により介入に参加する人々（遵守者）にとって介入の効果がどの程度かを推定します。

CARE 研究の結果の一部を記しておきます。手紙を受け取ったグループ（奨励群）は CARE を適用される場合が対照群に比べて 7.7%ポイント増加しました。ガス消費量は ITT による分析では 0.7%増加しました。LATE では CARE の適用を受けることによってガス消費量が約 9.1%増加したとしています。

がん検診や健康診断の大規模な RCT では、論文上では明記されていないのですが奨励デザインと似たような設計になっているものが多いです。

たとえば、大腸がんの内視鏡検査の効果を検証したヨーロッパの大規模な RCT では複数の国々の住民台帳を使って研究参加者を決めてランダムに 2 つのグループに分けて、1 つめのグループ（奨励群）にだけ内視鏡検査の案内状が送られました[12]。もう 1 つのグループ（対照群）は何も送られず、研究に参加していることも知らされませんでした（ただしノルウェーだけは対照群から研究参加の同意を取得)。この研究の分析は ITT で行われたのですが、大腸がんの発症リスクが 18%減少し（リスク比:0.82、95%信頼区間: 0.70～0.93)、大腸がんによる死亡には有意な効果がありませんでした(リスク比:0.90、95%信頼区間: 0.64～1.16)。この RCT では検査への参加率が約 4 割にとどまったため研究結果の妥当性に疑問が呈されました。

このような疑問を踏まえて、LATE の提唱者でノーベル経済学賞の受賞者であるアングリストが、医療関係の RCT であまり使われていない LATE を利用することによって、がん検診を実際に受けることの効果を明らかにすべきことを主張しています[13]。

---

[12] Bretthauer et al. (2022)
[13] Angrist and Hull (2023)

RCT の応用編２（クラスターRCT）

RCT に求められることとして、介入を受けた人々から介入を受けない人々が影響を受けないことがあります[14]。たとえば、学校の教室の中で介入を受ける人々と受けない人々を分けると、介入を受けた人々との接触を通じて介入を受けない人々が影響を受ける可能性があり、これは避ける必要があります。また、現実的に見て、同じ教室や学校内で異なるプログラムを生徒毎に提供するのは難しい場合もあります。

こうした課題に対応する上で、ランダム化の単位を教室や学校にすることが１つの解決策になります。これはクラスターRCT と呼ばれ、個々の人や企業ではなく、彼らが所属する組織（学校、教室、地域など）をランダム化することによって、介入効果を検証します。

クラスターRCT の例として、マスク着用が新型コロナ感染症を予防する効果があるかどうかを検証した大規模な研究があります[15]。バングラデシュの 600 の村に住む約 34 万人が研究対象となり、介入群（マスク着用を推奨するとともに無料のマスクを提供）となる村と対照群（何もしない）となる村に分けて、介入群の住民のコロナ感染割合が対照群と比べて低いかどうかを検証しました。対照群となった村ではマスクの着用割合が 13.3%で介入群では 42.3%でした。介入後に新型コロナ感染症の症状を有するに至った人々は対照群では 8.6%で介入群では 7.6%で、統計学的に有意な差がありました。仮にほとんど全ての住民がマスクを着用すればマスクの効果は更に大きくなるとこの研究の著者は推測しています。

なぜ RCT は重要か？

EBM や EBPM において RCT が重要な最大の理由は RCT が示すエビデンスが最も厳密であるためですが、それ以外にも RCT にはいくつかのメリットがあります。その一つは RCT がシンプルで透明性が高いことです[16]。

後述する RDD や DID といった分析手法は経済学や統計学の専門的知識がない人には手を出しにくいです。これに対して RCT は単純にサイコロを振って全体の集団を２つに分けて（偶数が出れば A グループ、奇数が出れば B グループ）、A グループには何かをしてもらい、B グループには待ってもらう、あるいは A グループと B グループには別なことをしてもらう、というそれだけのことで介入の効果検証が可能になります。

---

[14] Imbens and Rubin (2015)
[15] Abaluck et al. (2022)
[16] Gueron (2017)

また、RCT では結果の解釈が容易です。上記のビタミン D サプリの継続摂取の RCT では介入群の 12927 名のうち骨折したのは 769 名で、対照群の 12944 名では 782 名で、13 名の違いしかありません。難しい統計的な分析をしなくても、グループ分けがどの程度違った結果を生んだかが見るだけでわかります。RCT でも統計学的な分析は必要ですが、容易な場合が多いです。

RCT では段取りが重要です。人集めが大変だったり、アウトカムの測定が難しかったりしますし、医療などの RCT では本人から研究参加への同意を得ることや倫理審査委員会の承認取得などの手続きが求められますし、あらかじめ定められた手順通りに物事を進めていく必要があります。

その一方で、RCT を回していく中心人物には、統計学や計量経済学などの高度な知識は求められません。RCT でも統計学の専門家の参加は必要ですが、全体のプレーヤーの 1 人にとどまります。ただし、段取りを間違えないことの重要性は強調しすぎることはないです。たとえば、いったん行ったランダム化は崩さないのが鉄則です[17]。事前の設計を間違えると全てが水の泡になりますので、専門家を交えた慎重な設計が必要になります。

引用文献

Abaluck, J., Kwong, L. H., Styczynski, A., Haque, A., Kabir, M. A., Bates-Jefferys, E., . . . Mobarak, A. M. (2022). "Impact of community masking on COVID-19: A cluster-randomized trial in Bangladesh," *Science,* 375(6577), eabi9069.

Angrist, J. D., & Hull, P. (2023). "Instrumental variables methods reconcile intention-to-screen effects across pragmatic cancer screening trials," *Proceedings of the National Academy of Sciences,* 120(51), e2311556120.

Bretthauer, M., Løberg, M., Wieszczy, P., Kalager, M., Emilsson, L., Garborg, K., . . . Kaminski, M. F. (2022). "Effect of Colonoscopy Screening on Risks of Colorectal Cancer and Related Death," *New England Journal of Medicine,* 387(17), 1547-1556.

Gueron, J. M. (2017). Chapter 2 - The Politics and Practice of Social Experiments: Seeds of a Revolution. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of Economic Field Experiments* (Vol. 1, pp. 27-69): North-

[17] ランダム化が崩れることを contamination（日本語では「汚染」）と呼びます。対照群に割り振られた人々が介入を受けてしまうのが典型例です。

Holland.

Hahn, R. W., & Metcalfe, R. D. (2021). "Efficiency and equity impacts of energy subsidies," *American Economic Review,* 111(5), 1658-1688.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. New York, NY, US: Cambridge University Press.

LeBoff, M. S., Chou, S. H., Ratliff, K. A., Cook, N. R., Khurana, B., Kim, E., . . . Manson, J. E. (2022). "Supplemental Vitamin D and Incident Fractures in Midlife and Older Adults," *The New England journal of medicine,* 387(4), 299-309.

McCormack, D., Mai, X., & Chen, Y. (2017). "Determinants of vitamin D supplement use in Canadians," *Public Health Nutr,* 20(10), 1768-1774.

Sato, D., Sekizawa, Y., Sutoh, C., Hirano, Y., Okawa, S., Hirose, M., . . . Shimizu, E. (2022). "Effectiveness of Unguided Internet-Based Cognitive Behavioral Therapy and the Three Good Things Exercise for Insomnia: 3-Arm Randomized Controlled Trial," *J Med Internet Res,* 24(2), e28747.

Seligman, M. E. P., Steen, T. A., Park, N., & Peterson, C. (2005). "Positive psychology progress: empirical validation of interventions," *American Psychologist,* 60(5), 410-421.

What Works Centre for Local Economic Growth. (2016). "Evidence Review 4: Access to Finance. Updated June 2016."

リー，アンドリュー／上原裕美子訳. (2020). *『RCT 大全：ランダム化比較試験は世界をどう変えたのか』*：みすず書房.

加納悟・浅子和美. (1998). *『入門　経済のための統計学第２版』*：日本評論社.