無作為化臨床試験の考え方

高木廣文（新潟大学医学部保健学科）

１．無作為化臨床試験の基本的デザイン

これまで，科学的研究デザインとしてコホート研究，ケース・コントロール研究などについて説明した。しかし，実際にはよく知られているように，ＥＢＭの枠組みでは，無作為化臨床試験によって得られた研究結果が，最も科学的証拠として価値あるものとされている。

無作為化臨床試験では，対象者は比較すべき処理（治療法や薬剤など）を無作為（ランダム）に割当てられる。実験的デザインには，いくつかの細かな相違があるものが存在するが，その基本は対象の特定処理へのランダムな割当にある。このような無作為化を施したグループは，統計学的分析を適用して比較するのに，極めて優れたデータを与えてくれる。

それでは，なぜ対象者の処理への無作為割当を行う研究デザインが，最も科学的といえるのだろうか。この疑問は当然だと思うのだが，案外とこの点に関して説明をしている文書は多くない。この疑問に答えるために，今回はどのようにすれば治療効果や疾患発生に関する科学的な因果推論が行えるのかを説明する。そして因果推論のためには，どのような研究デザインが必要とされるかを説明する。

２．因果推論とカウンター・ファクチュアル・モデル

画期的な新薬Ａが開発され，これを飲めば絶対に循環器疾患による死亡はないとしよう。開発側はいくらそう考えていても，現実に新薬Ａの効果があることを証明するには，どのようにすればよいだろうか。もし，私がＡ薬を飲み続けて，心疾患でも脳血管疾患でも死ななければ，その効果があったと言ってよいのだろうか。逆の場合，すなわち，私がＡ薬を飲み続けたにも関わらず，心疾患が原因で死んでしまえば，Ａ薬は私には効果がなかったことは明らかである。

循環器疾患で亡くならなかった場合，それがＡ薬の効果であるという結論に対しては，いくつかの反論ができるだろう。例えば，運動を規則的にしていたからであるとか，肉油脂の摂取を控えた食生活のおかげであるとか，もっともらしいことが考えられるだろう。

このような反論に対して，どのようにすればＡ薬の循環器疾患に対する効果を，科学的に主張することができるだろうか。

正しい因果推論は，Ａ薬を飲む私と，Ａ薬を飲まない私を長期間追跡し，循環器疾患を発症するかもしくは死亡するかを調べ，両者の結果を比べればよい。このようにすれば，観察対象である私はＡ薬の服用の有無以外は完全に同一の条件なので，他の原因を完全に排除できる。

このような研究が可能ならば，Ａ薬の服用の有無と循環器疾患の発症の有無により，

⑴ Ａ薬服用の有無に関わらず，循環器疾患が発症する。

⑵ Ａ薬を服用した場合は発症せず，服用しない場合には発症する。

⑶ Ａ薬を服用した場合は発症し，服用しない場合には発症しない。

⑷ Ａ薬服用の有無に関わらず発症しない。

という４通りの場合の一つが観察されるはずである。

⑴の場合は，Ａ薬の服用に関わらず発症したので，Ａ薬は循環器疾患の発症に何の影響も与えていないことになる。

(2)の場合は，A薬は循環器疾患の発症を確かに抑制している。

(3)の場合は，A薬の服用は逆に循環器疾患の発症の原因になる。

(4)の場合は，(1)と同様に，A薬の服用は循環器疾患の発症に関係がないことになる。

このように，A薬の循環器疾患に対する効果を正確に調べることが可能である。しかし，現実には「A薬を服用する私」と「A薬を服用しない私」を同時に観察するためには，ＳＦでお馴染みのパラレルワールド（多元世界）が存在し，しかもその観察ができなければ研究は不可能である。このため，このような因果推論モデルは「反事実的モデル，仮定法的モデル，カウンター・ファクチャル・モデル counterfactual model」などと呼ばれている[1,2,3]。

３．因果推論のための次善の策は？

同一人物に対して，相反する２つの処理を同時に施して，その結果を観測することはできない。したがって，どのような処理であれ，処理とその効果に関する正確な因果推論は，理論的には不可能である。しかし，それでは新薬の開発や，新しい治療法などの評価ができないことになる。現実には，「科学的」に新薬の評価を行っているようだが，どのような考え方にしたがっているのだろうか。

完全に正確な因果推論ができないのならば，完全ではないがほぼ満足できるような次善の方法を，我々は考えるべきだろう。それでは，A薬の循環器疾患に対する抑制効果の有無を調べるための次善の方法とは，どのような方法だろうか。

正しい因果推論のためには，処理をした対象の観察と共に，その処理以外の点で全てが同一の比較のための対象が必要であった。このような比較のための対象は，「対照，コントロール control」と一般に呼ばれている。

もしも，私に一緒に暮らしている一卵性双生児の兄弟がいて，A薬を服用していなければ，A薬の影響をかなり正しく調べられるのではなかろうか。一卵性の双生児ならば遺伝的には全く同一なので，循環器疾患を発症するか否かは，薬剤の服用の有無と他の環境要因や生活習慣の相違によるものと言えるだろう。一緒に暮らしていれば，生活環境もほとんど同一となり，比較のためにはもってこいだろう。しかし，残念ながら，私には一卵性双生児の兄弟はいないし，全ての研究を一卵性双生児の兄弟姉妹を対象にして行うわけにはいかないだろう。

カウンター・ファクチュアル・モデルでは，A薬の私に対する影響を調べるためには，理論的にはA薬の服用以外は完全に同一な私が必要である。しかし，現実世界には完全に条件を満たすようなコントロールは存在しないし，遺伝的に同一な一卵性双生児の兄弟も存在しない。しかたがないので，さらに次善の策を考えねばならない。我々ができることといえば，私によく似た人間を捜すことぐらいだろう。この場合，「よく似た」ということの定義が難しいが，A薬による循環器疾患抑制効果に影響を与えるような属性が類似していることが必要だろう。少なくとも性や年齢，喫煙，飲酒，などの傾向は同じ方がよいだろう。

このように考えることで，一般にコントロールとしてどのような属性を持つことが望ましいのかが，ほぼ定まってくる。

４．標本は多いほどよい

私に対するA薬の循環器疾患の抑制効果を調べるためには，正しくはカウンター・ファクチャル・モデルによる必要があった。それでは，Ｂさんに対してはどうだろうか。当然，この答えも，正しくはカウンター・ファクチャル・モデルによる必要がある。Ｃさんに対しても，Ｄさんに対しても，すべて同

様ということになる。

もしも，本当に多重世界があり，ある世界の私はA薬服用者で，他の世界の私は非服用者であったというカウンター・ファクチャル・モデルによる調査研究が行えたとしよう。その結果，A薬服用者の私は循環器疾患を発症しなかったが，非服用者の私は心疾患を発症したとしよう。この結果，A薬は私に関しては確かに循環器疾患の抑制効果が証明された。ところが，Bさんについての観察結果では，どちらの世界のBさんも共に肺癌で死亡してしまった。すなわち，Bさんの場合，A薬の服用が循環器疾患と関係があるとはいえないことになる。また，Cさんは，A薬服用の場合には胃癌で，A薬非服用の場合には脳血管疾患で死亡した。まったく，三者三様であった。

このように，各個人は遺伝的にも環境的にも違いあるため，ある要因と疾患の因果関係は個別的ものである。また，肺癌で死亡してしまえば心疾患で死亡することはない。当たり前のことであるが，このように，ヒトの生死に関しては競合的な疾患が多数存在している。したがって，私という１例のみを調べても，それは私個人にしか当てはまらないことになる。すなわち，ケース・スタディは一般性を持っていないことになる。ケース・スタディがEBMの枠組みでは，あまり重視されないのは，このような理由に基づくといってよいだろう。

それでは，どうすればA薬の循環器疾患抑制効果を，より一般的に普遍化して研究することができるのだろうか。

各個人については，A薬の結果は一つしかあり得ない。それが２人の場合には対立する結果になるかも知れないし，同一の結果が得られるかも知れない。３人，４人，...，とその人数が増えるほど，個人の遺伝的な変動や他の環境要因からの影響がならされていく。ある程度の規模の集団になれば，その集団がもつ一般的な傾向，例えばA薬と循環器疾患の関係を明らかにすることができるだろう。この意味から，研究対象は可能な限り多い方がよいことが理解できるだろう。

ある処理の因果推論を正しく調べるために，カウンター・ファクチャル・モデルでは，ある処理を受けた人々を観察するとともに，その処理を受けなかった場合についても，同様に観察する必要がある。明らかに現実世界では，同一対象が同時に処理を受けたり受けなかったりするという，二律背反の状況を観察することはできない。このため，次善の策として，適当なコントロールを設定することになる。

コントロールとしては，Aさんによく似たａさん，Bさんによく似たｂさん，．．．，のようにとにかく研究したい要因や処理以外はできるだけ全ての特性が一致している必要がある。この操作は，ケース・コントロール研究でのマッチングに相当する。

しかし，性，年齢を一致させることはできても，他の特性を，個々人について一致させるのはほとんど不可能に近い。ケース・コントロール研究では，ケースである各患者に対して，どのようにコントロールを設定するかが大きな問題であった。それでは，臨床試験では，どのようにすれば処理群と未処理群の各対象が上記の点を満足できるようになるのだろうか。

５．無作為化を行う

実際にそれぞれ数百人ずつのA薬服用者と非服用者の集団を設定し，長期間の追跡研究を行う場合を考えてみよう。正しい因果推論のための理想的コントロールは服用者自身が非服用者となることであるが，次善のコントロールは，A薬服用という点以外はすべての条件が完全に一致するようにマッチされている必要がある。

現実的な問題として，そのようなマッチングは可能だろうか。研究対象者が少なければ可能かも知れない。よく似た２人ずつのペアを探してきて，一方にある処理を行い，もう一方にはその処理を行わず，

経過を観察することは可能かも知れない。

それでは，現実問題として「よく似た」とは，どのような場合を意味するのだろうか。いろいろな解釈が可能かも知れないが，疾病の因果推論に焦点をあてて考えてみよう。これは前提条件として「疾病の発生」に差をもたらさないように，リスク因子に関して同一条件を有している個体同士ということであろう。

一般に，多くの疾患に影響を及ぼす因子として，性と年齢がある。その他の因子も考えると，喫煙，飲酒，食物摂取状況，肥満度など，極めて多数ある。どの程度まで，各要因を一致すべきかは，実際の研究を行う場合には極めて重要な問題である。

ペアで一致すべき要因が増えれば増えるほど，そのようなペアを設定することが困難になり，結局，研究対象者数が十分に得られなくなるだろう。それでは，どのようにすればA薬服用者と非服用者のようなペアを，多数得ることができるだろうか。残念ながら，ペア単位，言い換えれば個々人の対象にマッチするようなペア造りでは，ペア数を増やすのには限界がある。

カウンター・ファクチャル・モデルによる正確な因果推論を諦めて，次善の策であるよく似た個人のペアの集団を観察するのもかなり難しいことが分かった。それでは，さらに次善の策として，よく似た集団同士を比較研究することにすればよいだろう。

そうは言っても，先ほどの議論と同様によく似た集団を創るために，各集団の構成員を一人一人選定していたのでは，どのくらい時間が必要か分からない。

そこで，まず研究対象者となる人たちのグループを設定する。例えば，健康な50歳代，60歳代の男性1000人を集めるなどである。当然，職業や飲酒状況などは対象者内でも個人差がかなりあるだろう。そのように個人個人は異なっていても，その一つの集団をよく似た２つの集団に分けることは可能である。

我々が，各個人の年齢や飲酒状況，生活習慣などを参考にして，各対象者を２つのグループに振り分けることも可能である。しかし，そのようにすると他の要因，例えば遺伝的な要因や他の食習慣などに偏りが生じるかも知れない。

結局，よく似た集団をつくるためには，無作為に集団の構成員を２つに分けるのが無難な方法である。すなわち，コインを投げて，Ａさんはコインの表なので第１グループ，Ｂさんは裏なので第２グループ，Ｃさんは...，のように人為を排除した方法により，各個人の割当グループを決めていけばよい。

実際には，コインやくじ引きではなく，コンピュータで「乱数」と呼ばれる数値を発生させ，その値にしたがって処理方法（A薬の服用，非服用など）の割当を決めて行くのである。乱数は，数字の並び方や発生順序に一定の規則性がなく，しかも数値の出現頻度に偏りがないというランダム性という要求を満たすものである。

教科書的には，これで目出度く比較すべき２グループが設定でき，各グループに処理か未処理かなどを割付，あとは実際の研究を行えばよいことになっている。

６．無作為割当で完全に２グループは同等か？

それでは，無作為に集団を分割すると，分割後の２グループの特性が，集団として完全に一致するのだろうか。残念ながら，実際には無作為割付によっても２グループの特性，例えば年齢の平均値，が完全に一致するとは限らない。とくにグループの人数が数十人程度のように，それほど大きくない場合には，無作為に分割された集団だからといって，各グループ間の特性に差が全くないということはほとんどない。

それでは，何のための無作為割当かという問題が生じるかも知れない。実際，無作為化の意味は，そ

れを無限回繰り返した場合，ある特性，例えば年齢の一回ごとの平均値の全試行での平均値（期待値）が各グループで一致する，ということを保証するものである。すなわち，個々の無作為化による施行に差がないことを保証するのではなく，同様の施行全体で平均として一致することを保証するものである。

したがって，類似の無作為化臨床試験を複数集めて，その結果をまとめて解析すれば，それが最もよい方法ではないかと考えられるだろう。そのような方法は，「メタ解析」と呼ばれており，方法論的には上記の意味から最も優れたものと見なされている。ただし，メタ解析は，基本的に極めて類似した方法により行われた研究をまとめたものでなければならない。異質な研究を単に寄せ集めてメタ解析をしても，よい結果が得られる訳ではない。

無作為化臨床試験で得られたデータを統計学的に解析する場合，無作為割当で２グループに分けたのだから，年齢に差はないはずであるだとか，ある特性の差を検討するのに年齢調整などは必要ないという議論は，このように間違いということになる。

実際には標本数が大きくなるにつれ，ある特性に関して，偶然に２グループでの特性差が大きくなる確率は，だんだんと小さくなることが分かっている。

因果推論を行うためには，現実問題として，無作為化の方法以外に我々は今のところ良い方法を持ち合わせていない。

７．統計学の役割

結局，科学的に正しく因果推論を行うための研究デザインを考えた場合，到達した結論は対象者の無作為割当による比較研究ということになる。しかし，そのような無作為化の方法であっても，研究している処理以外の特性に関して，比較すべき２グループ間で完全にその差をなくすことはできない。

しかし，無作為化の良い点は，偶然により生じ特性差に関して，どの程度の確率でそのような差が生じるのかを統計学的に評価することができることである。さらに，多変量解析を用いれば，研究目的でない変数の影響を除いた推定を行うことも可能である。

研究対象の選定の段階で人為的に生じた差を，客観的に評価する方法はない。しかし，偶然によって生じるグループ間の相違は，統計学というEBMの道具によって，客観的に評価することが可能である。

無作為化臨床試験に関して，さらに詳しく知りたい読者は文献１を，また今回の内容をより簡単に解説したものに文献３があるので参照されたい。

参考文献

1)椿広計，藤田利治，佐藤俊哉　編：これからの臨床試験－医薬品の科学的評価－原理と方法，21-33，1999，朝倉書店（東京）

2）Rothman，KJ，Greenland，S：Modern Epidemiology，Second Edition，49-50，1998，Lippincott-Raven Pub.（Philadelphia）

3)高木廣文，林邦彦：コントロールや無作為化はなぜ必要なのか，EBNursing，2(4)，102-107，2002