

医療における統計学の役割

浜田知久馬

医療の適正使用のためには、その使用の根拠は科学的な証拠に基づく必要がある。clarity, comparability, generalizability の3つの目標を達成した研究デザインによる科学的な証拠の創造と、得られた結果の解析・評価の両面で、統計学はEBM実現のために決定的に重要な役割を果たす。本稿ではコレステロール低下作用のあるプラバスタチンという薬剤の長期投与によって、高脂血症患者に対する、心血管系疾患発症抑制効果の検証を目的に行われた大規模臨床研究であるMEGA studyの事例に基づいて、医療における統計学の役割について解説する。

キーワード：Evidence-Based Medicine, clarity, comparability, generalizability

1. はじめに

クスリを後ろから読むとリスクになる。このことが象徴するように医薬品を適正に使えば、人の生命を救ったり、QOL (Quality of life: 生活の質) を大幅に向上させることができるが、使い方を間違えると副作用が生じ、薬害をまねくことは、繰り返されてきた不幸な歴史が物語っている。医薬品を適正に使用するためには、有効性と安全性に関する正確な情報が必要である。

医療ではEBM (Evidence-Based Medicine: 証拠に基づいた医療) が定着した[1]。EBMを実現させるために、科学的に強い証拠が得られるように、臨床研究の効率的な実験計画を考案し、得られた結果を適切に評価するための統計解析手法が必要である。統計学はEBMを実現するために、臨床研究のデザインと解析の双方で、決定的に重要な役割を果たす。本稿では大規模臨床研究であるMEGA studyの事例に基づいて、医療における統計学の役割について解説する。MEGA studyは、コレステロール低下作用のあるプラバスタチンという薬剤の長期投与によって、日本人における動脈硬化性疾患の既往のない軽度から中等度の高脂血症患者を対象に、心血管系疾患発症抑制効果の検証を目的に行われたもので、我が国で行われた最大規模の臨床研究である。結果、食事療法群と比較して、薬剤併用群では有意に疾患発生率が低下し、欧州で最も権威のある医薬ジャーナルLANCETに公表

された[2]。

2. 臨床研究の模式図

図1は典型的な臨床研究の模式図を示している。研究を始める際には、まず研究計画書(protocol)を作成する。このときに、研究対象とする疾患を規定するが、患者集団全体(いわゆる想定する母集団(population))が直接の研究対象となるわけではない。臨床研究では患者の適格基準と除外基準を設定して、研究対象集団を定義する。通常は腎臓や肝臓に重篤な疾患を持つ患者や、妊婦、極端な高齢者は安全性あるいは評価が困難であるという理由で除かれることになる。このため、対象疾患を持つ患者全体の母集団と、研究対象集団は厳密には一致しない。また併用禁止薬が設定され、投与方法や投与期間が厳密に規定されるので、実地医療とは多少異なることになる。研究計画書を作成後、各医療機関のIRB (Institutional Review Board: 施設内倫理委員会)で審査を行い、契約をかわし、登録期間中に来院した患者を試験に組み入れることになる。これが実際の研究対象集団、統計学でいうところの標本(sample)になる。

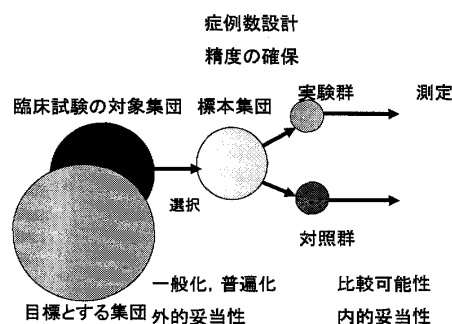


図1 臨床研究の模式図

はまだ ちくま
東京理科大学 工学部経営工学科
〒102-0073 千代田区九段北1-14-6

表1 医学研究における3つの目標

1	clarity (精度) を高くする。 Nを増やす。測定誤差を減らす
2	comparability (比較可能性) を保証する。 バイアスを減らす(無作為割付、二重盲検)
3	generalizability (一般化可能性) を保証する。 適格基準、除外基準の設定 交互作用解析、サブグループ解析

この対象者を通常、いくつかの治療法の異なる群に無作為に割付けて、薬剤を投与し、効果や安全性についての評価がなされる。一見して、随分単純なことを行っているようにみえるが、研究計画時に十分な考慮を行わないと、様々な落とし穴に陥る可能性がある。ここでは表1に示した臨床研究の質を高めるための、3つの目標について解説する。

3. clarity (精度) の確保

最初の目標は精度を高めて、なるべく明白な結論が出るようにすることである。clarity (精度) を高めるための工夫とよばれる。エンドポイント (臨床研究の評価指標) が血清コレステロール濃度等の臨床検査値である場合、測定の品質管理を行い、測定誤差を減少させることによって、研究の精度を高めることができる。例えば、臨床検査を施設ごとに行うと、測定方法等の違いによる施設間変動によって、バラツキが大きくなり、精度が低下してしまう恐れがある。このため、血清脂質等を主な評価指標とした臨床研究では、試験の精度を上げるために、血液のサンプルを中央で一括して、測定する場合もある。また精度を高めるために最も有効な方法は、症例数 N を増やすことである。Nが増えれば統計的な精度が高まり、より明白な結果が得られることになる。統計学的には N に比例して精度は増すが、Nを増やすと、費用も増し、また実験研究であるという倫理的な観点から、臨床研究の症例数は可能な限り少ないほうが望ましいので、必要な精度を保証する最低限の症例数で臨床研究を行う必要がある。このために統計学的な症例数設計が行われる。症例数設計に当たっては、前相の試験結果、海外で行われた臨床研究の結果等を参考にして、いくつかの必要条件を定めて、症例数を求めることになる。表2にMEGA studyの症例数設計の想定を示す。

MEGA studyの名の由来は、その大規模な症例数による。検証的な臨床研究では、あらかじめ、検出したい効果の大きさ、有意水準を定め、解析の精度を高

表2 MEGA studyの症例数設計

致死性/非致死性冠動脈疾患の発生率
5.6/1000 人年
食事療法群: 10%低下 併用群: 40%低下
ハザード比(発生率の比): 0.6/0.9=0.67
α : 10% (両側ログランク検定) 検出力: 80% (β エラー: 0.20)
脱落率: 20%
目標総症例数: 8000 例

めるために、見逃しの β エラーの確率が十分小さくなるように症例数を決定する。MEGA studyでは、致死性/非致死性冠動脈疾患の発生率を5.6/1,000人年と見積もり、薬剤効果の大きさとして、食事療法群に対する併用群の疾患の発生率の比(ハザード比)を0.67倍と予想した。両側ログランク検定を有意水準10%で行い、5年以上にわたる長期試験であることから、脱落率を20%と高めに想定すると、 β エラーを20%にするために必要な症例数は1群あたり約4,000人、2群をあわせた全体では8,000人となった。8,000人はかなり多いような印象を受けるが、疾患の発生率が5.6/1,000人年と低く、試験期間中に実際に疾患を起こす患者の割合もかなり低くなるため、精度を確保するために大規模な臨床研究が必要になった。

MEGA studyでは、試験を行った結果、無作為割付の対象となった患者数が計8,214例となり、解析対象例は食事療法群が3,966例、併用群は3,866例となった。事前の予想通り、比例ハザードモデルによる解析の結果、食事療法群に対する併用群の疾患の発生率の比が0.67倍となり、両側ログランク検定も $p=0.01$ で有意な結果が得られた[3]。

4. comparability (比較可能性) の保証

臨床研究では多くの場合、群間比較を行う。2番目の目標は、この群間比較の妥当性を保証することである。このための工夫を、comparability (比較可能性) を上げるといふ。また、研究のinternal validity (内的妥当性)を保証するともいわれる。研究の中での比較の妥当性を保証するため内的妥当性とよばれる。A薬群とB薬群があるとして、2つの群間で臨床成績に違いがあったとしたら、当該薬に起因したものであると断定したいわけである。このためには投与薬剤以外の要因が2群間で似通っている必要がある。2群間で性別・年齢・重症度等の背景因子の分布がそろっていれば、AとBの間で違いがあれば、それは薬剤の影響と断定することが可能になる。もし比較する群間で疾患の重症度の分布が異なっていれば、成績に違いが

あったとしても、治療の効果なのか重症度の違いなのか判別できない。比較可能性を保証するための常套手段が無作為割付である。無作為割付では、コインを投げて表が出ればA薬、裏が出ればB薬というように確率的に割付群を決定する。無作為割付を行うことにより、疾患の重症度や年齢等の背景因子の分布が、比較する2群間で平均的にそろることが期待できる。

また治療効果を評価する際に公平な比較を行うために、効果を評価する医師と評価される患者の両方に割付群をマスクする二重盲検が行われる。二重盲検を行わないと、薬剤を投与された患者は、治療に対する期待感から薬効を高め評価したり、不安感から副作用を多く報告する傾向が生じる。このため、通常は実薬から薬効の有効成分のみを除いた偽薬（プラセボ）を作成し、実薬と偽薬のどちらが投与されているか、医師と患者に判別できないように工夫する。無作為割付が患者の選択バイアスを防止するために行われるのに対し、二重盲検は評価バイアスを避けるために行われる。

MEGA studyでは研究のcomparabilityを保証するために、PROBE study (Prospective Randomized Open Blinded Endpoint Study: 前向き無作為化オープンエンドポイント盲検化試験)とよばれる研究デザインを採用した。このデザインでは無作為割付を行うが、二重盲検は行っておらず、医師も、患者も食事療法群と併用群のいずれに割付られたかを知っている。ただし、評価委員会において、試験のエンドポイント(致死性/非致死性冠動脈疾患の発生)に到達したかどうかを判断するときには、患者の割付られた群を隠して評価を行った。

MEGA studyでは、施設・性別・年齢で層別した置換ブロックを用いた層別割付法を用いた。サイズ6の置換ブロックを利用した。これは1つのブロックに、A, Bを2個ずつ含む6通りの配列(AABB, ABAB, ABBA, BAAB, BABA, BBAA)をコンピュータでランダムに発生させ、あらかじめ施設に薬剤を搬入し、施設ごとに、患者の性別・年齢によって層を決定し、患者の登録された順に配列にしたがい、Aであれば食事療法群、Bであれば併用療法群に割り付けるものである。MEGA studyは大規模な症例数を集めるため、多くの医療機関を対象とした多施設試験であり、施設間では対象患者の重症度の分布や標準治療の違いが存在する。このため局所的なバランスが保証されるブロック割付を採用した。各ブロックで

4例登録されれば、各群2例ずつ割り付けられ、バランスがとれる。ただし、ブロック割付を行うと、割付が予見できる場合が出てくる。1, 2例目の割付がAであれば、3例目はB以外ありえないためである。このため、MEGA studyでは登録センターを設け、割付情報の管理を行い、センターからの指示によって、割付方法を決定した。

このように無作為割付の方法を実際に適用する場合には、様々な工夫が必要であり、本来の目的を果たすために適切な手順が用いられているかを確認する必要がある。ある因子について積極的にバランスをとりたい場合は、層別割付や最小化法等の特別な方法が必要であり、目的にあった割付方法が用いられているかに、注意を払う必要がある[4]。実際に無作為割付がうまく機能し、群間で比較可能性のあることを保証するために、通常は、性別・年齢等の人口統計学的因子、疾患のステージ等の重要な予後因子の群ごとの分布を示し、群間で偏りがどうか検討した結果が示される。MEGA studyの論文でも、投与前の血清脂質等19項目の各群の背景因子の分布表を記述し、比較可能性のあることを示していた。

一方、MEGA studyで二重盲検を行わなかった理由は、薬剤の市販後の保険診療下で行われたため、前述の偽薬を用いることが困難であったこと、評価の際のバイアスは、評価委員会に対して盲検化することで回避できると考えたためである。投与薬剤であるプラバスタチンはコレステロールを20%程度に下げる効果があり、仮に二重盲検を行ったとしても、血清脂質を測定すれば、かなりの確率で、実際に受けた治療は判明してしまう。高脂血症患者に対して、長期に血清脂質の検査を禁止することは倫理的に不可能であり、実際上、二重盲検は不可能であった。

5. generalizability (一般化可能性) の保証

3番目の目標は、結果の一般化についてである。臨床研究の対象集団について、A薬群の方がB薬群より成績がよかった場合、単にこの研究対象者についてA薬の方がよいというだけでなく、結果をより一般化して、対象疾患を持つ患者全体の母集団に対して、薬剤の効果があると推測するのが研究の目的である。統計学では得られたデータを母集団からの標本とみなして、標本から母集団に対する推測を行う。この推測の妥当性をgeneralizability (一般化可能性) あるい

は external validity (外的妥当性) と呼ぶ。内部妥当性が研究の中での比較の妥当性を問題するのに対し、外的妥当性では研究結果を外に出すときの妥当性を問題にする。この一般化可能性を保証するために最も有効な手段は、無作為抽出 (random sampling) である。すなわち母集団から対象者を全て等しい確率でランダムに抽出することである。実際に無作為抽出は世論調査等で用いられ、無作為抽出された標本は母集団を代表することができる。無作為抽出を行うためには、母集団全体のデータベースを作成し、すべての患者に識別番号を付ける必要があるが、医薬研究では母集団を把握することは困難なので、無作為抽出が行われることはない。ランダム化という用語は、無作為抽出と無作為割付の2つの意味で用いられることがあるが、前者は一般化可能性を保証するための方法、後者は比較可能性を保証するための手段であり、役割は全く異なるので注意されたい。臨床研究の対象施設には、一般診療所は除かれ、大学病院等が選択されることが多いので、疾患の重症度等の分布が母集団全体とは異なっており、標本は対象とする母集団からランダムに抽出されたとみなすことは、一般的にはできない。それでは医薬研究で得られた結果を一般化するためには、どのような検討が必要だろうか。一般化可能性を検討する上では、研究デザインにおいて適格基準と除外基準をどのように設定したかが重要である。市販前の臨床研究の対象は狭い集団に限定されるので、特に薬剤の安全性については市販後モニタリングにより実地医療の中で監視を続ける必要がある。MEGA study では、主な適格条件は年齢 40~70 歳、体重 40 kg 以上の男性および閉経後女性、総コレステロール値 220~270 mg/dl であり、主な除外条件は家族性高コレステロール血症、冠動脈疾患または脳卒中の既往であった。したがって、70 歳を超えるような高齢者には MEGA study の結果は一般化できない可能性がある。

これに対し解析段階で一般化可能性を検討する手段が、サブグループ解析と交互作用 (interaction) 解析である。

サブグループ解析とは、背景因子で層を作り、層ごとに薬剤効果を推定することである。表 3 に疾患の改善をエンドポイントした場合に、重症度で層別したサブグループ解析の仮想例を示した。全体では改善率は A 薬群 50% に対して B 薬群では 60% であり B 薬には 10% の上乗せ効果があるが、この数字を一般化することは可能だろうか。重症度によって新薬の効果は

表 3 サブグループ解析の結果
(一般化可能性がない場合)

重症度	改善率(A薬)	改善率(B薬)	改善率の差
軽症	70%	70%	0%
中症	50%	60%	10%
重症	30%	50%	20%
全体	50%	60%	10%

表 4 サブグループ解析の結果
(一般化可能性がある場合)

重症度	改善率(A薬)	改善率(B薬)	改善率の差
軽症	70%	80%	10%
中症	50%	60%	10%
重症	30%	40%	10%
全体	50%	60%	10%

大きく異なる。軽症では上乗せ効果が 0% であるが、重症では 20% とかなり大きな効果がある。このように層ごとで薬剤効果の大きさが異なる場合を、重症度と薬剤に交互作用があるという。したがって全体効果の大きさは、重症度の分布に依存して大きく異なる。軽症に偏った集団では、全体の改善率の差は 10% を下回るし、重症に偏った集団では 10% を上回ることになる。これに対して表 4 のような場合はどうか。どの重症度でも一様に 10% の改善率の上乗せがあり、重症度と薬剤に交互作用は認められない。この状況では、重症度の分布がどのように変わっても薬剤効果は 10% の上乗せと一般化できることになる。対象とする母集団と臨床研究の対象集団では重症度の分布が異なるかもしれない。しかしながらサブグループ解析を行って、どの層でも同様の効果があれば、薬剤効果の大きさは母集団に一般化できることになる。背景因子で様々な層を構成して、サブグループ間で効果が一樣で交互作用がないことを示すことが一般化可能性を示唆することになる。

MEGA study の論文では、投与前血清脂質等の計 11 の背景因子でサブグループ解析を行い、サブグループごとに群間の致死性/非致死性冠動脈疾患の発生率の比を示し、また発生率の比にサブグループ間で違いがあるか、サブグループ×薬剤群の交互作用の検定結果を記述し、一般化可能性を検討した結果を示していた。多くの背景因子で、サブグループ間で薬剤効果に顕著な差はなく一般化可能性が示唆されたが、投与前 LDL コレステロール (悪玉コレステロール) につ

いては、低いサブグループでは疾患の発生率の比が0.90倍と効果が弱いのにに対し、高いサブグループでは0.54倍と顕著な効果が得られた。サブグループ間の交互作用の検定の結果も $p=0.11$ と、5%水準に近い結果が得られ、投与前LDLコレステロールが低い患者については、薬物療法を行う必要がないのかもしれない。

6. 終わりに

医療の適正使用のためには、その使用の根拠は科学的な証拠に基づく必要がある。clarity, comparability, generalizabilityの3つの目標を達成した研究デザインによる科学的な証拠の創造と、得られた結果の解析・評価の両面で、統計学はEBM実現のために決定的に重要な役割を果たす。MEGA studyではclarityを確保するために、多施設で大規模臨床研究を行い、疾患の発症については、専門医から構成される評価委員会で行った。comparabilityを保証するために、層別置換ブロック割付を用い、また評価委員会では盲検下で疾患の発症の有無を判定した。generalizabilityを検討するためにサブグループ解析、交互作用解析を行った。その研究結果は、十分、科学的証拠能力が高く、欧州で最も権威のある医薬ジャーナルLANCETに公表された。

本稿では紙面の制約上、医療のための統計手法の詳細については割愛した。現在では臨床研究の統計の考え方については、国際的なガイドラインとしてまとめられている[5]。方法論の詳細については成書を参照

されたい[6][7][8]。

参考文献

- [1] 名郷直樹, 「EBM実践ワークブックーよりよい治療をめざして」, 南江堂, (1999).
- [2] H. Nakamura, K. Arakawa, H. Itakura, A. Kitabatake, Y. Goto, T. Toyota, N. Nakaya, S. Nishimoto, M. Muranaka, A. Yamamoto, K. Mizuno and Y. Ohashi, "Primary prevention of cardiovascular disease with pravastatin in Japan, a prospective randomised controlled trial," *Lancet*, 368 (2006), 1155-1163.
- [3] 大橋靖雄, 浜田知久馬, 「生存時間解析」, 東大出版会, (1995).
- [4] A. Hagino, C. Hamada, I. Yoshimura, Y. Ohashi, J. Sakamoto and H. Nakazato, "Statistical comparison of random allocation methods in cancer clinical trials," *Controlled Clinical Trials*, 25-26 (2004), 572-584.
- [5] 厚生省医薬安全局, 「臨床試験のための統計的原則」. 医薬審第1047号, (1998). (<http://www.nihs.go.jp/dig/ich/eindex.html>よりpdfファイルとしてダウンロード可)
- [6] 浜田知久馬, 「学会・論文発表のための統計学—統計パッケージを誤用しないために」, 真興交易医書出版部, (1999).
- [7] 大橋靖雄, 荒川義弘, 「臨床試験の進め方」, 南江堂, (2006).
- [8] 上坂浩之, 「臨床試験の計画と解析」, 朝倉書店, (2006).