



## 医療統計学リテラシー(5)

## 正しい比較とは？

## ～リンゴとミカンを比べていないだろうか？～



大阪市立大学大学院医学研究科

医療統計学教室

新谷 歩

しんたに あゆみ

▶大阪市立大学大学院医学研究科医療管理医学講座医療統計学教授 ▶PhD(米国Yale大学)  
 ▶国立病院機構本部総合研究センター生物統計室長, 東海大学客員教授, 名古屋第一赤十字統計アドバイザー, 藤田医科大学統計アドバイザー, 米国Vanderbilt大学客員教授, 厚生労働省厚生科学審議会臨床研究部会委員・患者申出療養評価会議評価員 ▶1996年Yale大学修士課程修了, 2000年同大学博士課程修了, 01年Vanderbilt大学医療統計学部講師, 07年同准教授, 14年大阪大学大学院医学系研究科臨床統計疫学寄附講座教授, 16年より現職 ▶1968年生まれ, 兵庫県出身 ▶著書: 今日から使える医療統計, みんなの医療統計, あなたの臨床研究応援します ▶主な専門: ICUにおけるせん妄研究, 糖尿病, リウマチ, 癌, 感染症, 腎臓病など多分野にわたる臨床データの統計解析. 臨床研究ジャーナルに300篇以上に上る論文を執筆. 最近では米国国立衛生研究所(NIH)推奨の電子データ集積システムREDCapの日本への普及に尽力 ▶受賞: 2013年Vanderbilt大学医学部ティーチングアワード

新型コロナウイルス感染症の治療薬など, 新たに開発された治療薬の効果を調べるには, 研究に参加した被験者を, 投薬群とプラセボ群などの非投薬群にランダムに割り付ける無作為化比較試験が必要である。コインを投げて表が出たら投薬群, 裏が出たらプラセボ群というようにランダムに割り付けることで, 比較群間の被験者の年齢や性別などといった背景を揃えることができる。背景が揃うと, 比較群間のデータを直接比較することが可能になる。

しかしこの無作為化試験のような実験的な研究を行うには, 被験者が組み入れにくい, 費用がかかるなど施行に際してハードルが高く, 多くの臨床研究は無作為化比較試験でなく, 実臨床の患者データを直接用いることが多い。実臨床では, 治療された群はより重篤な人が多く入っているなど, 比較群間の背景が揃わないことが多い。

例えば新型コロナの治療薬を飲んだ群と飲まなかった群を比べる場合, 基礎疾患を持った高齢者のほうがより薬を飲んでおり, 同時に重症化率も高い場合, 治療をした群のほうがしない群に比べ, 重症化率が高くなり, 薬がまるで効かなかったかのようなおかしな結果になる。薬が効かなかったのではなく, もともと重症化しやすい人たちに投薬されていただけのことであるが, このように背景のズレを

考慮に入れず比較してしまうと薬の効果を間違っ  
て評価してしまう。このような比較群間の特性のズレを無視して比較し, 間違っ  
た結果を出してしまうことを, 英語では「Comparing apples to oranges (リンゴとミカンを比べる)」と呼び, 専門用語では「Confounding (交絡)」という。

表1の例は心臓病ハイリスクの患者を研究開始時にアスピリンを使用していたかどうかで群分けし, アスピリンを使うことで死亡率が下がるかを調べた観察研究において, 比較群間の背景を比べたものである。

この研究で得られたアスピリン投薬群と非投薬群の死亡リスクの差を数値化したハザード比は1.08と, アスピリンを用いると, 用いない群より1.08倍

表1 Gumらの観察研究における背景比較(全症例)

	アスピリン 使用群 (2,310人)	アスピリン 非使用群 (3,864人)	p値
平均年齢 (標準偏差)	62 (11)	56 (12)	<0.001
男性	77%	56%	<0.001
糖尿病歴	17%	11%	<0.001
高血圧	53%	41%	<0.001
冠動脈疾患歴	70%	20%	<0.001

死亡リスクが高くなることを示した。このエビデンスをもとに、果たしてアスピリンには効果がないと決定づけてよいだろうか。アスピリンを投薬された群は平均年齢が6歳高く、男性の割合も多く、冠動脈疾患の既往歴に関しては70%と、アスピリン非使用群の20%を大きく上回っている。つまり、アスピリンが効かなかったのではなく、よりリスクの高い人にアスピリンが投薬されたにすぎない。これは明らかに「交絡」が起こっている。

交絡の調整には多変量回帰モデルを用いた解析が使われる。回帰モデルとは我々が中学1年生で習う、 $Y=a+bX$ という式であり、Yにあたるものが臨床アウトカム、この場合は死亡を表す変数である。Xにあたるものが説明変数、この場合はアスピリンの有無を表す変数がXに相当する。この回帰モデルに、年齢・性別・重症度のような調整したい背景因子を加えることで、数学的に比較群のズレを数値化し、そのズレがアウトカムに及ぼす影響を取り除くことができる。Gumの研究では26個の背景因子を多変量解析で調整し、結果としてアスピリンの効果を示すハザード比は0.67 (P=0.002)。背景のズレを調整した結果、アスピリンの効果が示された。

◎観察研究のデータから背景のそろった比較群を抽出する傾向スコアマッチング

最近では、電子カルテやレセプトデータ等、何百万人というビッグデータを臨床研究に用いられるようになってきた。ビッグデータ解析などでよく使われる背景調整の手法に、傾向スコアによるマッチングがある。背景の揃っていない観察研究のデータから背景の揃った被験者のみを抜き出し、まるで無作為化されたようなデータを作り出す手法である。傾向スコアとは、複数の背景因子を一つの連続スコアに集約したもので、傾向スコアが同じ被験者をアスピリン使用群と非使用群からそれぞれ選びマッチングし、マッチされた症例のみを解析に用いる。背景が揃ったら、最終解析は無作為化研究のようにアウトカムを直接比較することができる。表2にGumの観察研究データからマッチングされた被験者のみを抜き出し、比較群の背景を示す。表1と異なり、マッチング後の背景はまるで無作為化したか

表2 Gumらによる観察研究において傾向スコアでマッチングした被験者における背景比較

	アスピリン 使用群 (1,350人)	アスピリン 非使用群 (1,350人)	p 値
平均年齢 (標準偏差)	60(11)	61(11)	0.16
男性	70% (951)	72% (974)	0.33
糖尿病歴	15% (203)	15% (207)	0.83
高血圧	50% (679)	49% (659)	0.79
冠動脈疾患歴	48% (652)	49% (659)	0.79

のように揃っているのが分かる。

傾向スコアの「傾向」とは、背景情報から推測したそれぞれの被験者が研究対象となる治療を受けている確率である。例えば、より重篤な人のほうが治療を受けやすいといった場合は、傾向スコアは重篤度スコアに類似する。治療あり・なし群など2群を比較する場合の傾向スコアは通常、2値のロジスティック回帰モデルを用いて行う。無料の統計ソフトのEZRでもロジスティック回帰モデルの画面に傾向スコアを作成するというオプションがあり、またマッチングまで自動的に行うことができる。

下のQRコードよりEZRを用いた多変量回帰モデルと傾向スコアマッチングの使い方のビデオを紹介する。



多変量解析ロジスティック回帰モデルの使い方



多変量 Cox 比例ハザードモデルの使い方



傾向スコアマッチングの使い方

参考文献

Gum PA, et al. : Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: A propensity analysis. JAMA, 286 (10) : 1187 ~ 1194, 2001.

今月号で、大阪市立大学大学院医学研究科・医療統計学教室教授の新谷歩氏による連載は終了となります。次号からは、同じく「人間と科学」というシリーズタイトルの下に、京都外国語大学・京都外国語短期大学副学長の黒住淳人氏による連載が始まる予定です。