大学院医学研究科シリーズ

間違いやすい統計解析

千葉康敬

近畿大学医学部附属病院臨床研究センター

1．はじめに

まずは次の新聞記事（某有名全国紙2001年４月某日の朝刊：一部改変）を読んでほしい．

> Ｃ型肝炎治療のインターフェロン
> 効果，遺伝子が左右
> ○○大教授ら解明
>
> Ｃ型肝炎の治療薬インターフェロンの効果が，患者の遺伝子のわずかな違いで左右されることを○○大の△△教授らが見つけた．この薬が効きやすい人と，そうでない人がいて，遺伝情報の個人差が影響すると考えられていたが，具体的な遺伝子がわかったのは初めてという．治療方針の判断に使えそうだと期待される．
>
> この薬は免疫を刺激してＣ型肝炎ウイルスが感染した細胞ごとウイルスを壊すとされる．
>
> △△教授らは，免疫反応に関連した酵素をつくる複数の遺伝子を研究．治療効果の差は，LMP7という遺伝子のタイプで生じていた．タイプは２種類あり，塩基配列が１カ所だけ違っている．
>
> インターフェロン治療によってウイルスが消えた49人と，消えなかった126人を対象にこの遺伝子を調べた．ウイルスが消えた人の約16％が持っているタイプは，消えなかった人では約８％と少なかった．
>
> ウイルスの量が比較的少ない場合，このタイプの人だと８割の人でウイルスが消えたのに，そうでない場合は５割だった．
>
> この遺伝子がつくる酵素は，ウイルス感染を免疫細胞に知らせるのに関係する．タイプの違いが働きの差になり，治療効果に影響するようだ．
>
> インターフェロンはうつ状態になるなど副作用がある．△△教授は「治療効果の予測や適切な治療期間の決定などに応用できる」と話している．

なんら疑問に感じることなく，さらっと読んで納得してしまってはいないだろうか？

実は，この記事の中には，本来なら言えるはずのないことを言っているところがある．それが何なのか，なぜそれが言えないのか，を正しく理解できるようになることは，医学論文を読む者全員にとって非常に重要なことである．

本稿では，臨床研究の論文を読む際の１つの視点を提供しつつ，生物統計学の落とし穴ともいうべき事柄について議論する．まず，２節と３節でここに示した記事を適切に読み解くための統計的事項について説明し，その上で，４節でおかしな点を解説する．５節では他の記事を例に挙げ，６節と７節でその記事を適切に読み解くための統計的事項について説明し，８節でおかしな点を解説する．最後に９節で本稿のまとめを行う．

統計と言っても，テクニカルな話は一切ないし，特別な事前知識も必要としない．肩の力を抜いて気軽に読んで頂ければと思う．なお，本稿の内容は拙著[1]によるところが大きいことを断っておく．

2．前向き研究と後ろ向き研究

さて，前節の新聞記事であるが，何がどう「よろしくない」かがわかるためには，研究デザインについて正しく知っておく必要がある．

飲酒と心疾患との関係を調べることを例に考えてみよう．この関係を調べる１つの方法として，はじめに，飲酒ありグループと飲酒なしグループのように，原因となるものでグループ分けし，その後数年間追跡調査をして，その間に心疾患を発症したかしなかったか，というように，対象としているイベントが発生したかどうかを調べる方法がある．このように，調べようとしている方向が原因（現在）から結果（未来）へと前向きになっている研究デザインのことを前向き研究と呼ぶ．

それに対して，はじめに，心疾患ありの人と心疾患なしの人のように，対象としている疾患に現在罹っている人と罹っていない人を集めてきて，それから，飲酒の有無のように，原因となるものがあったかなかったかを，過去の資料やインタビューに基づいて調べる方法もある．前向き研究とは違って，現在から過去に遡ってデータを収集しようというわけである．このように，調べようとしている方向が結果（現在）から原因（過去）へと後ろ向きになっている研究デザインのことを後ろ向き研究と呼ぶ．

研究デザインに関する解説は本誌にもある[2]ので，詳しくはそちらを参照してほしい．ここで問題としたいのは，前向き研究か後ろ向き研究かで統計解析の方法が異なるということである．このことを見るために，「オッズ比」という指標を考えてみよう．

3．「オッズ比」という指標

3.1．前向き研究でのオッズ比とリスク比

ある都市で，心疾患発生のリスクが高いと考えられる人を全員集めてきて，飲酒ありと飲酒なしの2つのグループに分けて，5年間追跡調査したとする．その結果，表1が得られたとしよう．

このデータを使ってオッズ比を計算してみよう．前向き研究では，オッズ比は曝露（今の場合，飲酒）ありグループのオッズと曝露なしグループのオッズの比と定義される．ここでのオッズは，イベント（今の場合，心疾患）が起きるリスクとイベントが起きないリスクの比のことである．

表1のデータでは，飲酒ありグループでのイベント（心疾患）が起きるリスクが800/10,000，イベントが起きないリスクが9,200/10,000なので，飲酒ありグループのオッズは

$$\frac{800}{10,000}\Big/\frac{9,200}{10,000}=\frac{800}{9,200}$$

となる．同様に，飲酒なしグループのオッズは

$$\frac{400}{10,000}\Big/\frac{9,600}{10,000}=\frac{400}{9,600}$$

となる．よって，オッズ比は

$$\frac{800}{9,200}\Big/\frac{400}{9,600}=2.09 \qquad (1)$$

となる．このように計算されるオッズ比は，実は，そのままでは解釈不可能な指標なのである．前に述べたが，オッズは，イベントが起きるリスクとイベントが起きないリスクの比のことである．定義として字面を追うことはできても，その意味するところは不明である．意味不明なもの同士の比であるオッズ比も当然意味不明で，解釈不可能なのである．

だったらなぜオッズ比などという解釈不可能な指標を計算するのだろうか？

この疑問を解消するために，まず，オッズ比とリスク比の関連についてみてみよう．表1のデータでリスク比を計算すると，

$$\frac{800}{10,000}\Big/\frac{400}{10,000}=2.00 \qquad (2)$$

となる．飲酒ありグループでの心疾患発生リスクは飲酒なしグループの心疾患発生リスクよりも2倍高かった，ということである．

オッズ比が2.09だったので，オッズ比とリスク比が近い値をとっていることがわかる．オッズ比の計算(1)とリスク比の計算(2)を見比べると，「800」と「400」は同じように分子に配置されているので，分母の，「9,200」と「10,000」，「9,600」と「10,000」が近い値をとっていれば，オッズ比の値はリスク比の値に近くなる．これを表1で見てみると，飲酒あり，飲酒なしの各グループで，「合計人数」と「心疾患なし」の人数が近ければ，オッズ比の値はリスク比の値に近くなることがわかる．言い換えると，「心疾患あり」の人数が少なければ，オッズ比の値はリスク比の値に近くなる．つまり，発生が稀な疾患では，オッズ比はリスク比の近似値となるのである．

3.2．後ろ向き研究でのオッズ比

ここで例に挙げているデータでは，全対象者数が20,000人であった．20,000人もの多くの人を調査するのは大変だし，「心疾患なし」の人が「心疾患あり」の人の10倍以上もいてバランスも悪いので，「心疾患なし」の人を1/10の1,880人だけランダムサンプリング（無作為抽出）して後ろ向き研究を行ったとしよう．そうすると，ランダムサンプリングしているので，飲酒ありの人数も1/10，飲酒なしの人数も1/10となって，表2の結果が得られることが期待され

表1 全例を対象とした飲酒と心疾患の仮想的な研究結果

飲酒	心疾患		合計
	あり	なし	
あり	800	9,200	10,000
なし	400	9,600	10,000
合計	1,200	18,800	20,000

表2 飲酒と心疾患の関係を調べる仮想的な後ろ向き研究の結果

飲酒	心疾患		合計
	あり	なし	
あり	800	920	1,720
なし	400	960	1,360
合計	1,200	1,880	3,080

る．

表２のデータを使ってオッズ比を計算してみよう．後ろ向き研究でのオッズは，前向き研究でのオッズと違って，曝露を受けた割合と曝露を受けなかった割合の比として定義される．

今の場合，心疾患ありグループでは，飲酒ありの割合が800/1,200，飲酒なしの割合が400/1,200となる．よって，心疾患ありグループのオッズは

$$\frac{800}{1,200}\Big/\frac{400}{1,200}=\frac{800}{400}$$

となる．同様に，心疾患なしグループのオッズは

$$\frac{920}{1,880}\Big/\frac{960}{1,880}=\frac{920}{960}$$

となる．だから，オッズ比は

$$\frac{800}{400}\Big/\frac{920}{960}=2.09$$

となる．もともとの全対象者20,000人でのオッズ比とまったく同じ値になった．これは偶然ではなくて必ず成立することである．

これまで述べてきたことから，

・後ろ向き研究で定義されるオッズ比の値は前向き研究で定義されるオッズ比の値に等しい
・前向き研究でのオッズ比は，発生が稀な疾患であれば，リスク比の近似値となる

ことがわかった．これらのことから，発生が稀な疾患であれば，後ろ向き研究でのオッズ比もリスク比の近似値になると言える．

しかし，所詮近似は近似である．それ自身では解釈不可能なオッズ比よりも，できるものならはじめからリスク比を計算した方が良いのである．では，後ろ向き研究でリスク比を計算することを考えてみよう．表２のデータを用いて，飲酒ありの人たちと飲酒なしの人たちでリスクを計算してみると

・飲酒ありの人たちでのリスク＝800/1,720＝0.47
・飲酒なしの人たちでのリスク＝400/1,360＝0.29

となる．ところが，もともとの全対象者20,000人でのリスクは

・飲酒ありの人たちでのリスク＝800/10,000＝0.08
・飲酒なしの人たちでのリスク＝400/10,000＝0.04

である．ぜんぜん違う値になっている．このことからわかるように，後ろ向き研究でリスクを計算すると，間違った値が算出されてしまうことになる．したがって，後ろ向き研究ではリスクを計算してはいけないことになる．リスクが間違った値になるので，リスクの比であるリスク比やリスクの差であるリスク差も間違った値として算出される．後ろ向き研究ではリスク比やリスク差も計算してはいけないのである．一方で，オッズ比はもともとの全対象者20,000人で計算した場合と同じだった．だから，特別な情報を用いて特殊な統計解析[3]をしない限り，後ろ向き研究ではオッズ比を計算しなければならないのである．

ちなみに，前向き研究では，ランダムサンプリングしたデータでも適切にリスクを計算できる．例えば，飲酒ありグループの人を1/10の1,000人だけランダムサンプリングしたとすると，飲酒ありの人数も1/10の80人になることが期待される．よって，飲酒ありの人たちでのリスクは80/1,000＝0.08となり，もともとの全対象者20,000人でのリスクと同じになる．

4．研究デザインの重要性

ここまでに述べてきたことを踏まえた上で，もう一度１節で紹介した新聞記事をみてみよう．この研究では，ある特定の遺伝子のタイプとウイルス消失の有無の関係を調べている．「ある特定の遺伝子のタイプ」が原因で「ウイルス消失の有無」が結果である．

４段落目を見てみると，「インターフェロン治療によってウイルスが消えた49人と，消えなかった126人を対象にこの遺伝子を調べた．」との記載がある．このことから，この研究は，調べようとしている方向が結果（ウイルスの消失）から原因（遺伝子）へと後ろ向きになっている後ろ向き研究であることがわかる．

ところが，その次の段落を読んでみると，「ウイルスの量が比較的少ない場合，このタイプの人だと８割の人でウイルスが消えたのに，そうでない場合は５割だった．」との記載がある．これより，遺伝子のタイプごとにウイルスが消えた割合，つまり，リスクを計算していることがわかる．3.2節で述べたように，後ろ向き研究ではリスクを計算してはいけないのに，である．これが本来言えるはずのないことである．

この例では，特別な情報を用いて特殊な統計解析をしない限り，オッズ比を計算するしかない．ウイルスの消失が稀なイベントでなければ，その計算結果は解釈不可能となるにもかかわらず，である．

臨床研究においては，データをどのように解析するかはもちろん重要であるが，それにも増して，データをどのように取得するか，つまり，研究デザインをどうするかが非常に重要となる．研究デザインによって統計解析の方法が（ある程度）決まる．用いる統計解析の方法によって導き出せる結果や結論が決まる．臨床研究においては，研究デザインを考えることなしに正しい結論を得ることはできないの

である．

5．仮想的なランダム化臨床試験の例

ここまでは研究デザインについて議論してきた．ここからは臨床研究の結果から正しい結論を得ることについて議論したい．まずは，次の仮想的なランダム化臨床試験の結果報告[4]（一部改変）を読んでほしい．

新しい小児用中心静脈カテーテルの有用性評価

目的：新しい小児用中心静脈カテーテルは，従来品に比べて挿入の成功割合が高いかどうかを調べた．

方法：100人の小児を2グループにランダム割り付けし，グループAで新しい小児用中心静脈カテーテルAを，グループBで従来からある小児用中心静脈カテーテルBを挿入した．

1回の試みでカテーテル挿入が成功した割合を比較した．また，挿入時間と挿入の容易度も比較した．

挿入成功割合はフィッシャーの直接法，挿入時間はt検定，挿入の容易度はカイ2乗検定で比較した．$p<0.05$で有意差あり，と判定した．

結果：1回の試みでカテーテルを挿入することが可能であったのは，グループAでは50例中47例，グループBでは50例中44例であった．これらの2グループ間での挿入成功割合に有意な差はなかった．

挿入時間は，グループAの方がグループBに比べ有意に短かった（$p<0.05$，差の95%信頼区間（秒）：-2.8，-1.2）．またカテーテルの挿入が容易と判断されたのは，グループAでは40例，グループBでは36例で有意な差はなかった．

結語：新しい小児用中心静脈カテーテルAは，従来からある小児用中心静脈カテーテルBに比べ，より短時間に挿入できたため，カテーテルAはBに比べてより有用である．

この文章については，何の疑問も感じることなく読み終えることはないだろう．なぜなら，この研究の目的が新旧カテーテルの「挿入成功割合」を比較することであるのにもかかわらず，結論がカテーテルの「挿入時間」に基づいてなされているからである．目的に応じた測定項目である「主要評価項目」以外の測定項目である「副次的評価項目」で結論付けがなされてはいけないのである．挿入成功割合で有意差がなくても，挿入時間で有意差があるのだから，カテーテルAはBに比べてより有用であると結論付けても良いのでは？　と思うかもしれないが，いけないのである．なぜだろうか？

これがわかるためには，統計的仮説検定の（基本的な）考え方を知っておかなければならない．

誤解するといけないので一応付け加えておくが，この文章は，引用文献[4]中でもよろしくない例として紹介されている．

6．統計的仮説検定の方法

6.1．統計的仮説検定の原理

はじめに統計的仮説検定の原理について述べる．一見関係なさそうであるが，次のことについて考えてみよう．

2014年6月某日，大阪狭山市で殺人事件が発生した．近畿大学医学部附属病院で働くA氏は殺人の容疑をかけられてしまった．容疑を晴らすためにはどうすればよいだろうか？

一番確実なのは，アリバイがあることを証明することである．推定殺人時刻前後に，出張などで大阪狭山市にいなかったことを証明すればよい．

この「アリバイを証明する」ということについて，少し理屈っぽく考えてみると次のようになる．

まず，A氏が殺人を犯したという仮説を立ててみよう．この仮説が正しければ，殺人が特殊な遠隔殺人などでない限り，推定殺人時刻前後にA氏は大阪狭山市にいたはずである．よって，その時刻に大阪狭山市にいなかったことが証明できれば，殺人を犯していないことが証明できることになる．つまり，「殺人を犯していない」ことを証明するために，わざわざ逆の「殺人を犯した」という仮説を立てて，それを否定することにより，殺人を犯していないことを証明するのである．

注意しなければならないのは，もし仮にアリバイがなかったとしても，それが殺人の証拠にはならないということである．その時刻に大阪狭山市にはたくさんの人がいたはずで，アリバイがなかった人もたくさんいるはずである．唯一言えることは，「犯人でないとは言えない」ということのみである．

統計的仮説検定は，このアリバイ証明の原理を応用しているのである．

例えば，かぜ薬を飲むグループと飲まないグループを比較するランダム化研究では，「薬に効果があ

る」ことを証明したいはずである．このことを証明するために，わざわざ逆の「薬に効果がない」，つまり，「比較するグループの風邪が治った割合に違いがない」という仮説（帰無仮説）を立てて，それを否定しようというわけである．得られたデータで統計的仮説検定をして帰無仮説を否定できれば，「比較するグループの風邪が治った割合（リスク）に違いがある」と結論付けられる．しかし，アリバイがなかったからと言って即犯人だと断定できないように，帰無仮説が否定できなかったからといって，「比較するグループのリスクに違いがない」（薬の効果はない）とは言えない．「比較するグループのリスクに違いがあるとは言えない」としか結論付けられないのである．

では，得られたデータからどうやって帰無仮説を否定できるかどうかを判断すればよいのだろうか？

6.2．たまたまの可能性を考える

仮想的なランダム化研究のシンプルな例を見ながら考えてみよう．薬を飲むか飲まないかによって翌朝に風邪が治るかどうかを調べるランダム化臨床研究を行ったとする．そうしたら，200人の人が参加してくれて，表3の結果が得られた．

リスク差を計算すると，

$$70/100-60/100=0.10$$

となる．

さて，この0.10というリスク差だが，本当に薬に効果があって出てきた数値なのだろうか？

もしかすると，本当は薬の効果がなくてリスク差が0のはずなのに，たまたまの偶然の影響によって出てきた数値なのかもしれない．ランダム割り付けしたとしても，すべての要因が比較するグループ間でピッタリ等しくなることはなく，たまたま何かしらの要因が偏る可能性がある．

そこで，リスク差の値が0であると仮定して，先ほどの結果がたまたまの偶然の影響によって生じてしまった可能性がどのくらいあるのかを調べてみよう．「帰無仮説」という言葉を使うと，帰無仮説（リスク差＝0）が正しいと考えたときに，たまたまの偶然の影響によって，データから推定されたリスク差以上に極端な値（0.10以上の値）が生じてしまう可能性がどのくらいあるのか，を調べてみようということである．

表3　仮想的なランダム化研究の結果

グループ	風邪		合計
	治った	治らなかった	
薬を飲む	70	30	100
薬を飲まない	60	40	100

便宜上，薬を飲むグループでも飲まないグループでも，ちょうど間をとって，

$$(70+60)/(100+100)=65\%$$

の割合で風邪が治るはずだと考えよう．薬を飲むグループ100人のうち65人は風邪が治り，同じように，薬を飲まないグループ100人のうち65人は風邪が治るはずだと考えるわけである．このとき，リスク差は

$$65/100-65/100=0$$

になるはずである．

しかし，薬の効果がなかったとしても，たまたまの偶然の影響によって，2つのグループ間でリスクに差が生じてしまうことがある．この偶然の影響によるリスク差のブレ幅を，コンピュータシミュレーションでみてみることにしよう．手順は以下の通りである．

① 薬を飲むグループの100人が確率65％で1，確率35％で0が出るように乱数を発生させる．
② 薬を飲まないグループの100人が確率65％で1，確率35％で0が出るように乱数を発生させる．
③ ①と②で，「1」を「風邪が治った」，「0」を「治らなかった」と置き換えて，グループごとに風邪が治る人が何％いるかを計算し，そこからリスク差を計算する．
④ ①～③の作業を1000回繰り返す．

図1はこのシミュレーションの結果を表している．

横軸はリスク差を示し，縦軸は1000回中のその頻度を示している．例えば，横軸の「0.00」のところは，リスク差が－0.01以上0.01未満だった回数が，1000回中119回あった，ということを意味する．リスク差が必ずしもちょうど0にならないのは，コイン

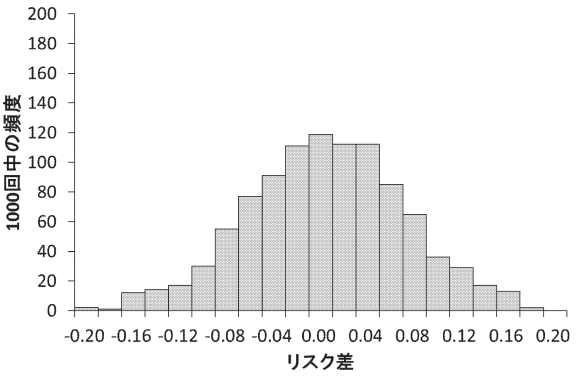

図1　200人でのシミュレーション結果

トスを100回して表が出る回数がちょうど50回になるとは限らないのと同じ原理である．

リスク差がデータから計算されたリスク差である0.10以上になったのは，1000回中62回だった．本当は差がないはずなのに，たまたまの偶然の影響によってリスク差が0.10以上になってしまう可能性が6.2%ある，と解釈することができる．ちなみに，反対側に帰無仮説から同じくらい離れる，リスク差が−0.10以下になった回数も，1000回中62回(6.2%)だった．

6.3．p値と有意差の有無

この6.2%という数値こそが(片側)p値なのである．つまり，p値というのは，帰無仮説が正しい(比較するグループのリスクに違いがない)と考えたときに，たまたまの偶然の影響によって，データから推定されたリスク差以上に極端なリスク差が計算される可能性のことなのである．一応付け加えておくが，リスク差ではなくて，リスク比や他の効果の指標であっても同様である．また，実際にはシミュレーションではなくて理論式を用いてp値を計算する．よって，正確に計算したp値は若干異なる．

先ほどの例で言うと，

・片側p値＝6.2%
・両側p値＝6.2%＋6.2%＝12.4%

となる．この片側p値は，本当は差がないはずなのに，たまたまの偶然の影響によってリスク差が0.10以上になってしまう可能性のことである．両側p値は，本当は差がないはずなのに，たまたまの偶然の影響によってリスク差が0.10以上または−0.10以下になってしまう可能性のことである．

もしもp値がとても小さければ，「リスク差が0だと仮定したときに，たまたまの偶然の影響によってリスク差が0.10以上または−0.10以下と計算されてしまう」可能性がとても低い，と考えられる．そうだとすれば，「現実のデータで可能性の低いことがたまたま起こった」と考えるよりは，「リスク差が0だという仮定(帰無仮説)が間違っている」，すなわち，「リスク差は0ではない」(薬の効果は0ではない)と考える方が自然である．

では，p値がどのくらい小さければ「リスク差は0ではない」と考えればよいのだろうか？

明確な論理的根拠はないが，医学領域では，慣例的に，しばしば両側で5%(片側2.5%)という基準が用いられている．両側p値が5%よりも小さければ，たまたまの可能性は考えにくい，要するに「リスク差は0ではない」と判断するのである．本当のリスク差は0なのに，たまたまの偶然の影響によって誤ってリスク差は0ではないと判断してしまう可能性が5%あることになるけれども，それぐらいは許容しましょう，というわけである．この便宜的に設けた基準値が有意水準である．

この例では，有意水準両側5%で判断するということは，「リスク差が0だと仮定したときに，たまたまの偶然の影響によってリスク差が0.10以上または−0.10以下と計算される可能性」(p値)が5%未満なら(たまたまの可能性が5%未満だったら)「リスク差が0だという仮定(帰無仮説)が間違っている」と判断しましょう，ということになる．しばしば，

・p値＜有意水準　なら　「有意差あり」
・p値≧有意水準　なら　「有意差なし」

という言い方をする．先ほどの例を見ると，

両側p値＝12.4%≧5%

なので，「有意差なし」ということになる．ただし，注意してほしいのは，6.1節で述べたように，「リスク差は0である(リスクに違いはない)」とは結論付けられなくて，あくまでも「リスクに違いがあるとは言えない」のである．

7．検定すれば良いというものではない

7.1．人数によって変わるp値

6.1節で行ったのと同じシミュレーションを，表4に示した倍の400人でしてみよう．

リスク差は，この400人の場合でも，140/200−120/200＝0.10である．

6.1節の200人では，本当のリスク差を65/100−65/100＝0と仮定してコンピュータシミュレーションした結果，図1が得られた．同じようにして，400人の場合でも，本当のリスク差を130/200−130/200＝0と仮定してコンピュータシミュレーションすると，図2が得られた．

200人の場合(図1)よりも，400人の場合(図2)の方が，リスク差が0のあたりに集中しているのがわかる．

リスク差が0.10以上になった1000回中の頻度等も比較してみよう．表5を見てほしい．

リスク差が0.10以上になった回数は，200人の場合には1000回中62回あったのに対して，400人の場合には1000回中20回しかなかった．有意水準両側5%で統計的仮説検定をしてみると，200人の場合は「両側

表4　人数を倍にした場合の結果

グループ	風邪		合計
	治った	治らなかった	
薬を飲む	70×2＝140	30×2＝60	100×2＝200
薬を飲まない	60×2＝120	40×2＝80	100×2＝200

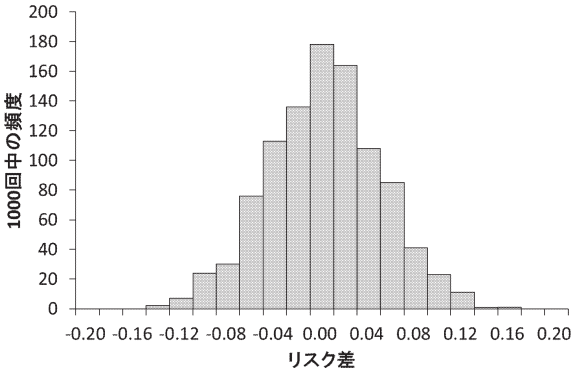

図2　400人でのシミュレーション結果

表5　200人の場合と400人の場合のｐ値の比較

	リスク差（1000回中の頻度）		
	0.10以上	−0.10以下	合計
200人の場合	62回	62回	124回
400人の場合	20回	19回	39回

ｐ値＝12.4％≧５％」だから「有意差なし」だが，400人の場合は「両側ｐ値＝3.9％＜５％」だから「有意差あり」となる．つまり，リスク差の値が同じであっても，ランダム割り付けされる人数が違うだけで，「有意差なし」「有意差あり」の結果が変わってしまうことがある．極端な話，すごくたくさんの人でランダム化研究をすると，たとえリスク差が0.00001だったとしても「有意差あり」ということになってしまうことがある．医学的にはまったく何の意味もないような差であっても，人数が多いだけで「有意差あり」となってしまうことがあるのだ．逆に言えば，医学的にはとても意味のあるような差であっても，人数が少ないだけで「有意差なし」となってしまうこともある．

では，統計的仮説検定をすることに何の意味があるのだろうか？　ｐ値を計算することに何の意味があるのだろうか？

実は，ただ漠然とｐ値を計算したり統計的仮説検定をしたりすることには大きな意義はないのである．大きな意義があるのは，医学的に意味のある差があるときには「有意差あり」，医学的に意味のない差のときには「有意差なし」となるように，あらかじめ研究に参加してもらう人数を正しく計算し，つまり症例数設計を行って，その通りの人数で研究を行ったときのみなのである．

症例数設計の原理や計算は幾分厄介である．文献[1,5]を参照してほしい．ただし，文献５は中級者向けである．

7.2．複数の測定項目での検定

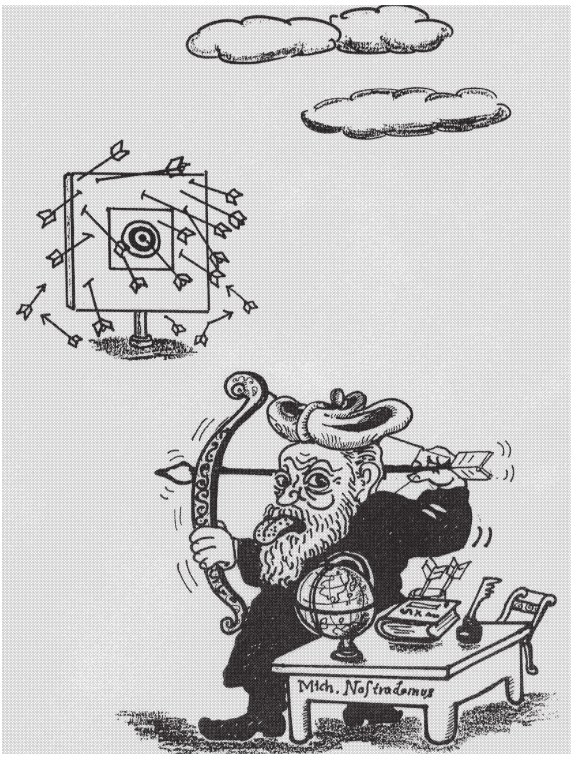

図3　ノストラダムスの大予言[6]

５節に示した仮想的なランダム化臨床試験の例では，「挿入成功割合」「挿入時間」「挿入の容易度」の３つの測定項目について統計的仮説検定を行っていた．このように，複数の測定項目について統計的仮説検定をすることを考えてみよう．

6.2節でみたように，有意水準両側５％で統計的仮説検定をするということは，本当は差がないにもかかわらず，たまたまの偶然の影響で間違って差があると判断してしまう可能性が５％ある，ということである．複数の測定項目で統計的仮説検定をすると，本当は差がないのに間違って差があると判断してしまう可能性が，測定項目の数の分だけ多くなってしまうのである．下手な鉄砲も数撃ちゃ当たるし，ノストラダムスの予言も，たくさんあるうちのいくつかは当たるのである（図３）[6]．

がむしゃらになんでもかんでもｐ値を計算したり統計的仮説検定をしたりすることには大きな意義はないのである．大きな意義があるのは，原則として，研究の目的に見合った主要評価項目についてのもの１つだけなのである．

8．研究結果の解釈における注意点

ここまでに述べてきたことを踏まえた上で，もう一度５節で紹介した仮想的なランダム化臨床試験の結果報告をみてみよう．この試験では，新しい小児

用中心静脈カテーテルの有用性を調べている．主要評価項目は挿入成功割合である．

この試験では，３つの測定項目（挿入成功割合，挿入時間，挿入の容易度）について統計的仮説検定をし，このうち，挿入時間でのみ有意差があった．7.2節で述べたように，複数の測定項目で統計的仮説検定をすると，本当は差がないのに間違って差があると判断してしまう可能性が増える．挿入時間についても，本当はグループ間で差がないのに，たまたま有意差のある結果になってしまっただけのかもしれない．

ここで「方法」の項に注目してみよう．「100人の小児を２グループにランダム割り付けし，…」と記載されている．7.1節で述べたように，この100人が，主要評価項目である挿入成功割合について，医学的に意味のある差があるときには「有意差あり」，医学的に意味のない差のときには「有意差なし」となるように決めた症例数であれば，この統計的仮説検定の結果には大きな意義がある．「有意差なし」という結果には大きな意義があって，「カテーテルＡはＢに比べて有用であるとは言えない」と強く主張できるのである．

そうではなくて，適当に集めた100人で試験を実施していたのであれば，医学的に意味のある差があるときには「有意差あり」，医学的に意味のない差のときには「有意差なし」となっていない可能性が高い．この場合には，統計的仮説検定の結果には大きな意義はない．得られた結果の信頼度は低く，仮に有意差があったとしても，何も強く主張できないことになる．ましてや，症例数設計を行っていない副次的評価項目（挿入時間，挿入の容易度）の統計的仮説検定の結果に大きな意義がある可能性は極めて低い．副次的評価項目の統計的仮説検定の結果については，その信頼性は低いと判断すべきである．基本的には，参考程度の情報だと考えるのが妥当である．

9．おわりに

本稿で述べたことからわかるように，臨床研究の論文を読む際には，研究の目的，その目的に見合った主要評価項目，研究デザイン，その研究に必要な症例数が明確に記載されているか否かに注意するとよい．これらの情報が明記されていないものは，いい加減に研究を行っていたり，何かやましいことがあってそれを隠していたりする可能性がある．信頼度の低い研究だと考えてほぼ間違いないだろう．

臨床研究の質を評価するポイントは他にもある．具体的なポイントについては紙面の都合上ここでは述べないが，CONSORT 声明[7] というものが非常に参考になる．CONSORT というのは，CONsolidated Standards Of Reporting Trials の略で，「臨床試験報告に関する統合基準」のことである．25項目からなるチェックリストがあり，そのチェックリストに基づいてチェックすれば良いわけである．

臨床研究を実施する際には，このチェックリストを満たすように計画を立てて，プロトコール（研究実施計画書）にまとめれば良いことになる．ただし，これは非常に大変な作業であって，言うほど簡単なものではない．具体的にプロトコールに記載する事項については，がん領域に特化しているが，良い文献[8] があるのでそちらを参照してほしい．本学倫理委員会のホームページからもプロトコールのテンプレート[9] をダウンロードすることができる．論文執筆時にも CONSORT チェックリストは役に立つであろう．

近年，impact factor の高い雑誌を中心に，生物統計学の専門家が査読者に入ることが増えてきた．ｐ値のとても小さい positive data であるといくら主張しても，その信頼性が低ければ採択される可能性が低くなってきているのである．

臨床研究を実施する医学研究者のみならず，その結果報告を見る医療関係者全員にとって，統計学は避けられないものとなってきている．生物統計学の基本的な考え方を身に付けることは，医療関係者にとってますます重要になっていくだろう．

謝　辞

執筆の機会を与えて頂いた免疫学教室宮澤正顯教授に感謝します．

文　献

1．千葉康敬：医療統計の基礎がギュッとつまった本（仮題）．総合医学社（2015年１月頃発刊予定）
2．伊木雅之（2012）人を対象にした研究デザイン．近畿大学医学雑誌37：203-210
3．佐藤俊哉（1992）ケース・コントロール研究再考．医学のあゆみ162：225-226
4．浅井　隆：いまさら誰にも聞けない医学統計の基礎のキソ３（研究の質を評価できるようになろう！）．アトムス
5．山口拓洋：サンプルサイズの設計．健康医療評価研究機構
6．浜田知久馬：学会・論文発表のための統計学．真興交易医書出版部
7．津谷喜一郎，元雄良治，中山建夫訳（2010）CONSORT 2010声明：ランダム化並行群間比較試験報告のための最新版ガイドライン．薬理と治療38：939-947
8．中村健一，福田治彦（2009）臨床試験プロトコールの書き方３．腫瘍内科３：357-364
9．近畿大学医学部附属病院臨床研究実施要項：http://www.med.kindai.ac.jp/rinri/files/youkou.pdf