

Neyman-Pearson 統計学から新しい臨床統計学へ： エビデンスよりは説明責任を

医学統計解析グループ（代表：前谷俊三）

天理よろづ相談所 医学研究所

患者中心の医療や治療選択のインフォームドコンセント、及び医療者の説明責任を要求する声が高まる中で、統計学にも変革が求められている。本稿では現行のNeyman-Pearson 統計学の問題点を指摘し、患者や社会が理解でき、治療の選択を助けるための新しい臨床統計学を提唱する。よい治療法を選択するための第一の基準は平均的治療効果の差（期待値基準）である。ただこの基準に沿って選んだ治療が他の治療よりも悪い結果をもたらすリスクが多少とも存在する。そこで第二の基準としてこの確率（治療選択を誤るリスク） P' を推定する必要がある。そのためには第三の仮説を設ける。これは帰無仮説や対立仮説と比べて現実に最も起こり得る仮説であり、「標本から推定したパラメータ（平均、分散）が真のパラメータ（母平均、母分散）に一致する」と仮定する。この仮説の下では P' 値は片側 P 値に近似する。これはまたSchwartzのいうエラー（第3種の過誤）であり、エラーやエラーと比べて、患者にとって第一の関心事である。個々の患者にとっての P' と集団にとっての P' とは峻別すべきである。期待値基準と前者の P' を組み合わせることで、より緻密な臨床決定ができる。なお生存分析においてはハザード比よりは、数理モデルから予測した平均余命の差が治療効果の評価に適している。これにより費用効果比も計算できる。新しい臨床統計学は証拠に基づく医療を超えて、患者や社会に対して説明責任を果たすものでなければならない。

キーワード：証拠に基づく医療（EBM）、医療者の説明責任、期待値基準、エラー、平均余命

はじめに

今医療は大きい変革の流れの中にある。それは最近の医療に関するキーワードにも

反映されている。例えば「エビデンスに基づく医療（evidence-based medicine, EBM）」や、「医療の質の評価」、「医療費の高騰と抑制」、「医療者の説明責任（professional accountability）」、「医療情報の開示、その透明性と対称性」、「患者中心の医療」、「患者

【別刷請求先】

〒632-8552 天理市三島町200
天理よろづ相談所 医学研究所
前谷俊三

の自己決定権とインフォームドコンセント」, などである。この時期において, 医療と深く関わりあってきた統計学も, 変革の流れに無縁ではいられない。それでは統計学において何が変わらねばならないのだろうか。現行の統計学は非専門家にとっては余りにも観念的で難解である。統計解析結果が患者や社会のわかる言葉で正確に伝えられ, 彼らが最良の医療を自ら選べるまでには程遠いといわねばならない。統計情報の非対称性は医療を受ける側と提供する側の間にみられるだけでなく, 医療関係者の中でも統計専門家と非専門家の間にも存在する。また非専門家側の理解不足や誤解は少なくない。例えばFaheyは健康政策に携わる人々に, 乳癌検診計画とトリハピリ計画のランダム化比較試験の結果を示し, どの程度この計画を支持するかアンケートをとった。その結果, 同じ結果を違った統計的指標で示すと, 支持率が大きく変わった。¹

既に1967年フランスのSchwartzは, 英米を中心とするNeyman-Pearson流の考え方が統計学の全てではなく, これとは別の実際的な考え方があることをpragmatic approachという言葉で表現し, 臨床統計学の進むべき一つの方向を示唆した。² われわれはこれまで本誌でも現行の統計学の問題点を指摘してきたが,³⁻⁵ 今回はできるだけ数式を避け卑近な例を挙げて, 治療法の選択において誤解をきたしやすい問題を論じる。それに基づき臨床統計学の新しい考え方を提唱したい。

治癒日数の群間比較(皮膚疾患に新薬を使うべきか)

一例として両側下肢を冒す皮膚疾患に対

してある新薬が効くことが示唆された。その新薬は局所のみ作用し, 全身的な影響はないものとする。そこで100人の患者を被検者として, 下肢のどちらか一侧のみにat randomに新薬を使用し, 反対側は対照として放置した。両側で治癒までに要した日数を数え, その差から新薬が対照と比べて, 治癒日数をどれだけ促進するかを調べた。図1上段はその結果をヒストグラムで図示した架空のデータである。平均すると新薬で治癒が3.4日促進されたが, 中には薬を使った方が治癒が遅れた例(治癒促進日数が負となる例)もあり, 症例によるばらつきが大きかった。果たして新薬は有効といえるだろうか。また使用に値するだろうか。

1. 患者にとってわかりやすい治療法の選択基準

もし新薬が無料であり, 副作用などの不都合もないとすれば, 新薬を使うかどうかの第一の判断基準は平均治癒日数が短縮することである。これは期待値基準⁶といわれ, 臨床決定分析において治療の選択の拠り所となる。⁷ しかし, もし新薬を入手するのに自費で1万円要するとすれば, それを承知で新薬を推奨してよいだろうか。新薬には余分の費用がかかったり, その他で不利な点があれば, 平均治癒促進日数だけでは新薬使用の是非を判断することは難しく, 治癒促進日数のばらつきも考慮しなければならない。

ここで図1の上段を見ると, ヒストグラムは0を境として色分けされている。0を分岐点としたのはこれを境にして, 新薬の価値が逆転するからである。これをcritical pointと呼ぶことにする。これより左側の黒く塗りつぶした部分の面積は43%である。

これは薬を使ったために、使わないよりも反って治癒が遅れた被検者が100人中43人いたことを示す。言い換えれば、これは「平均すれば得をする筈の新薬を使って、反って損をするリスク」である。Kahneman(2002年ノーベル経済学賞受賞者)によれば、絶対値が同じでも損をするほうが得をするよりも過大に評価され、損失を回避しようとする心理がみられるという。⁸ 新薬を使えば

使わない時と比べていつでも治癒日数が短縮する場合と、新薬で反って治癒が遅れる危険がある場合とでは、平均治癒促進日数が同じでも、新薬の値打ちに大きい違いが生じる。この意味で黒塗りの面積は期待値基準とは別の重要な判断規準となる。

2. 対応のある検定と新薬の価値

次に新薬の有効性を通常使用される統計的方法で調べればどうなるだろうか。この

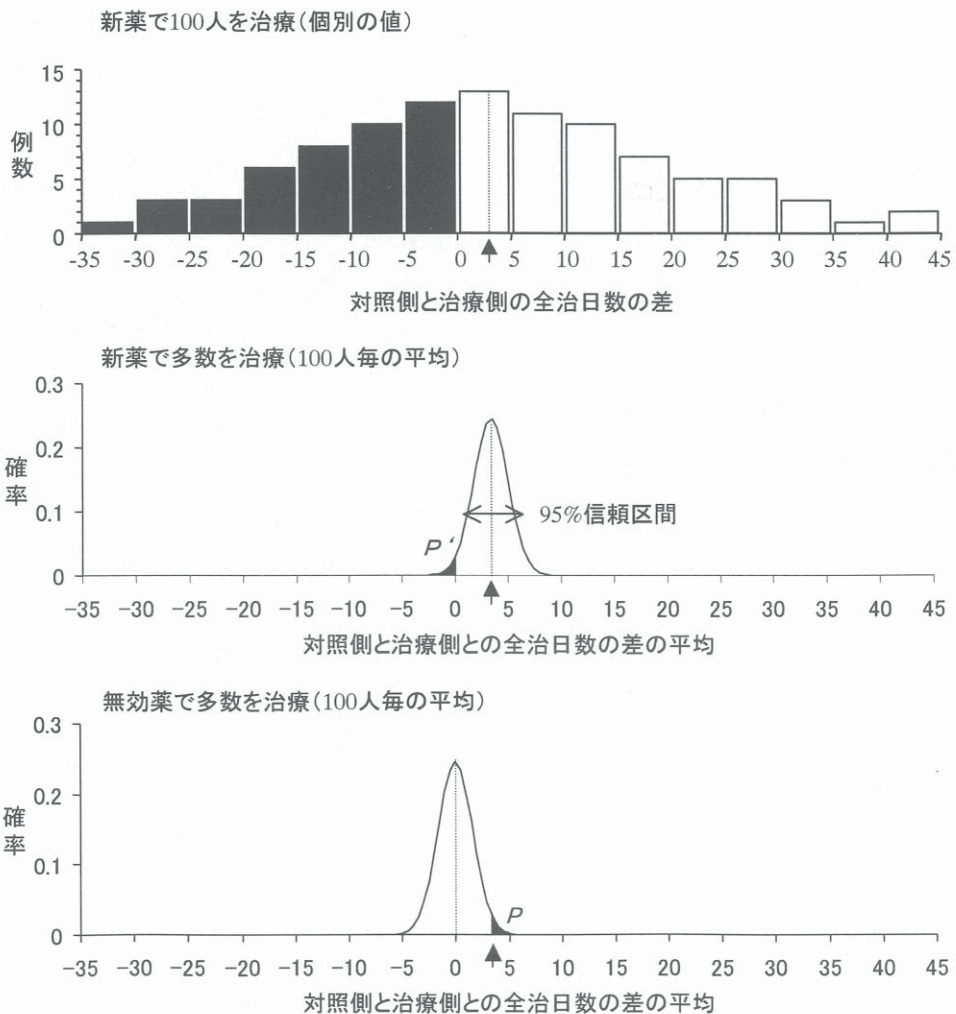


図1. 両側下肢皮膚疾患において一侧に新薬を使用し他側(対照)と全治日数を比較した臨床試験矢印は平均を表す。

例では同一被検者,同一条件下で新薬と対照との治癒日数を比較したので,使用すべき検定法はおなじみの「対応のある t 検定」である.その結果は $t=2.09$ (自由度99)となり, $P=0.04$ (両側検定)で,新薬と対照との間に統計学的に有意な差がみられた.一方,平均治癒促進日数の95%信頼区間は,0.17-6.63日と正の側に収まるので,新薬の有効性が再確認された.この信頼区間は図1の中段に示す.

中段には同時に平均治癒促進日数の分布も併記している.この分布は上段の観察データから理論的に導いた分布であり,その意味は以下の通りである.今回と同じように,新たに100人の被験者を使って何度も同様の試験を繰り返したとすると,個々の平均治癒促進日数は常に3.4日になるとは限らない.時には3.1日になったり,3.5日になることもある.個々の平均値(標本平均)は本当の平均値(母平均)の周りに分布する.その形は図1中段のような正規分布に近くなる.実はこの標本平均の分布を描くためにはある仮定を設けている.それは「実際に観察された平均治癒促進日数(3.4日)は母平均と一致する」である.もっと厳密に言えば「観測データ(標本)から推定したパラメータは母集団のパラメータと一致する」である.後述するように,これは極めて重要な仮説であり,従来の帰無仮説や対立仮説に対して第三の仮説と呼ぶことにする.この分布の0より左側で黒く塗りつぶした部分の面積を P' とすると,僅か2%である. P' を見れば新薬を使えば大抵の患者で治癒日数が短縮するような錯覚を覚える.上段と中段とではばらつきが大きく異なる理由は,標本平均の分布は個々の値の分布と比べて

著しく幅が狭くなっているためである(理論的には標準偏差は $1/\sqrt{\text{症例数}}=1/10$ に縮小する).個々の値の分布と平均の分布のうち,果たしてどちらが患者にとって有用な情報であろうか.EBMにおいては,図1中段に示すように平均値だけでなく,その信頼区間をエビデンスとして重視した⁹.しかしこれだけで患者は納得するだろうか.

3. 標本平均の分布は十分な判断基準となるか

ここでもうひとつわかりやすい例を挙げる.一流大学に入るためにある予備校を受験した生徒が,合格通知と共に50万円の入学金を請求された.そこで入学金を支払う前に,その予備校で勉強すれば一流大学に入れる可能性はどれだけかと問い合わせた.その回答は「その大学の合格ラインは700点であるが,我が校生徒で行った模擬試験の平均は855点である.95%信頼区間でいえば,730-980点であり,合格ラインを越えている」であった.これは詐欺ともとられかねない的外れの(irrelevant)情報である.提示すべき情報は全生徒の模擬試験点数の分布でなければならない.特に700点未満の生徒は何%いたかを示すべきであり,700点がここではcritical pointである.この情報を開示しなければ,50万円を払ってその予備校に入る価値があるかどうかは判断できない.生徒数さえ多ければ,たとえ合格ラインに達しない生徒が半数近くいたとしても,平均値の信頼区間は合格ラインをクリアすることもある.

4. 標本平均の分布の意味

ここで図1中段の平均治癒促進日数についてもう少し掘り下げて考えてみよう.第三の仮説の下では,薬で反って治癒が遅れるものの割合(黒塗りの部分)は,図1上段

の43%から中段では2%に減少する。その理由は100人の被験者の平均とは、個々の被験者の治癒促進日数をプールして平等に再配分したことと同じだからである。その結果、極端に大きい値や小さい値がなくなり、ばらつきも、新薬が裏目に出るリスクも共に減ることとなる。

もし再配分の対象が金銭であれば、被験者同士で取得した利益を返上して平等に再配分することはできる。しかし対象が治癒日数であれば、それは不可能である。実現できない平均値の分布を提示してどのような意味があるのだろうか。

平均値のもう一つの意味は、個人の損得ではなく、集団が得た利益、即ち、全員の損得の総和を表すという考え方である。但し平均値は総和そのものではなく、個人当たり換算した値である。集団が共有する利害も個人の損得も平均値にすれば変わりはない。しかしそのばらつきは、個人に比べて集団では相対的に小さくなり、しかも集団が大所帯になるほどその傾向は強くなる。

5. 平均値の群間比較とランダム化比較試験

上の例では一人一人の被検者で新薬と対照とが同時に比較できた。しかし一人の被検者に対して新薬か対照かどちらか一方だけしか試みられない場合の方が多い。その場合は新薬群と対照群とは別々の集団となり、多少とも群間に偏りが生じる。それをできるだけ減らすためにランダム化が行われる。両群の比較には、対応ある t 検定に代わって、対応のない t 検定が使用される。これが普通に行われているランダム化比較試験(RCT)であり、並行試験といわれる。このデザインではもはや図1上段に示すような個々の被験者の治癒促進日数は測れない。

中段に示すように新薬の効果の平均と、対照の効果の平均との比較、言い換えれば集団と集団との比較しかできない。

EBMは医療における最適の選択肢を迫する方法であるが、問題は誰にとって最適かである。EBMの提唱者の一人であるSackettはEBMの目的を個々の患者のケア - (the care of individual patient)に置いている。¹⁰しかし彼らが最も強いエビデンスをもたらすと考えるRCTでは、個々の患者の損得は評価できないのが普通である。通常のRCTは飽くまでも集団の利害を測る方法に過ぎない。上の例のように、「一人の患者が1万円を払って新薬を使う値打ちがあるか」という問いには、答えを出せないのが普通である。答えられることは、「被検者と同数の100人の患者集団に対して、新薬のために100万円(1万円×100)を投資する価値があるか」という問題に対してである。これは個々の患者よりは、医療政策者や保険者にとって関心のある問題であり、国や団体にとって有用な情報である。その1例を以下に示す。

6. P' 値は一定ではない

ある会社の従業員の中で丁度100人が上述の皮膚疾患に罹患し、全治するまで休業を要したとする。もし1人1日の休業で会社は1万円の損失を蒙るとすると、1万円の新薬で平均1日以上治癒が早まれば、新薬を使う価値がある。そこで図1中段のcritical pointを0日から1日に変えて P' を求めると、約7%となる。その意味するところは「新薬を使って会社が損をするリスクは約7%、つまり得をする確率は93%」ということになる。罹患患者数が100人より多い会社ではリスクは更に減るかもしれない。ところ

が罹患する従業員が僅か1人ならば話は別であり、個人単位での損得と同じになる。この場合のリスクの計算には図1上段のヒストグラムを利用すべきであり、新薬を使って反って損をする危険は43%以上になる。この危険は普通のRCTでは求められない。もしこのヒストグラムの分布が非対称で右裾が長ければ（治癒促進日数が異常に長い例があれば）、その危険は50%を越えることもある。

以上からわかるように P' 値には個人にとっての値や集団にとっての値があり、また critical value の選び方によっても異なる値をとる。

7. α , β , エラーと P' 値との関係

ここでもう一度図1の中段に戻って復習をしてみよう。100人という集団に新薬を使って治癒日数の総和を求め、これと対照との治癒日数の総和を比較した場合、既に述べた第三の仮説の下では、新薬を使った場合の総治癒日数が対照よりも反って長くなる確率はわずか2%である。つまり100人という集団にとっては、新薬を使って損をするリスクは、個人に新薬を使う場合と比べて著しく減少する。これが左裾の黒く塗りつぶした部分 (P') であり、Schwartz のいう エラー (第3種の過誤) をおかす確率である。² 但し新薬と対照とで治癒日数を厳密に測れば、両者の治癒日数が等しくなること (治癒促進日数=0.000...) はないと仮定している。エラーは エラー (第1種の過誤) や エラー (第2種の過誤) ほどは知られていないので、念のためこの三者を説明する。

今2つの治療A、Bの効果を比較したいとする(A、Bの一つは無治療でもよい)。エ

ラーとは、「A、Bの効果には差がないのに、一方が他方よりよく効く」と見誤ることである。エラーとは「A、Bの効果には差があるのに、差がない」と見誤ることである (厳密には エラーとは帰無仮説が正しいのにこれを否定することであり、エラーとは対立仮説が正しいのにこれを否定することである)。これに対して エラーとは「Aの方がBより効くのに、Bの方が効く」と見誤ることである。あるいはこの反対でもよく、要はA、Bの優劣を取り違えることである。元来エラーとはそれを犯すと何らかの損失を蒙るものである。以上のなかでどの誤りをすると損失が最も大きいかを考えると、それは エラーであることは明白である。一方、エラーとは、元々差のない治療の中で、一方が効くと思ひこむことであり、その結果として一方の治療をしたとしても、他方の治療をするのと患者に及ぼす効果は変わらないので、実害はない。エラーによる損失はこれより大きい、エラーによる損失よりは小さい!¹¹ とすれば第一にコントロールすべきは エラーではないだろうか。

ところが現行の Neyman-Pearson 流の統計学では エラーは無視して、エラーをできるだけ小さく抑えようとしている。例えば、RCTで予め必要症例数を算定する場合、エラーは5%以内の確率でしか発生しないようにコントロールしているが、エラーは通常10%乃至20%と、これより大目にみている。各エラーがどれだけの損失をもたらすかを考えた上で最適の対応をしているのだろうか。筆者はかつて Lancet でこの疑問を投げかけたが、これに対して納得のゆく説明は得られなかった。^{12,13} しかしそ

の後この疑問は次のように説明すればよいことがわかった。

8. P' 値と P 値

ここで図1の下段に目を転じてみよう。ここでは新薬ではなく、偽薬のような効果のない治療を100人に行った場合の平均治療促進日数の分布である。当然のことながらそれは0を中心とした分布をする。黒塗りの部分は平均治療促進日数3.4日(矢印)より右側の面積で0.02となる。これがおなじみの P 値(片側)である。これと図1中段の P' を比べると両者は同じ値となる³。その理由はそれらの値を決めるのは分布曲線のパラメータ(平均と分散)及びcritical valueであるが、 P' のcritical value(=0)が P の平均に、 P のCritical value(=3.4)が P' の平均と入れ替わっているだけで、他は同じだからである。従って P 値を小さくすればエラーも小さくなる。これによって始めて現行のエラーを小さく制御することの意味が納得できる。

それでは P' と P の値の誤差はどちらが大きいだろうか。この値を決めるのは分布曲線の分散と二つの数値0と3.4である。この中で3.4は標本によって変動するので誤差の原因となる。しかし P' と P の計算が共にこの3.4を使うので、誤差の程度に変わりはない。問題は分布曲線の分散の推定に両者とも標本分散を使用していることである。本来これは観察されたデータの下では最もあり得べき推定値である(第三の仮説)。しかし P 値の基になる帰無仮説において同じ値を使ってよいだろうか。帰無仮説は「新薬が毒にも薬にもならない」ことを意味している。しかし、もし新薬が活性をもつ場合には、創治療を促進する例もあるが、逆

に有害反応により創治療を遅らせる例もあり得る。これは根治手術という治療では寿命を延ばす効果もあれば、時には無治療よりももっと寿命を縮める効果があるのと同様である。もしこの二つの効果が重なれば、無効薬と比べて図1上段の治療促進日数のばらつきが広がる。そうでなくても有効な治療は治療促進日数の平均だけでなく、分布の幅や形を変えることは十分考えられる。とすれば帰無仮説の下での分布に同じ分散を当てはめることには聊か問題がある。

なお上述の話は集団にとっての P' 値のことであり、個人にとっての P' 値ではない。この意味では P' 値は P 値よりはもっと広い概念といえる。

我々の臨床統計学とNeyman-Pearson流の統計学との違い

P と P' とが同じ値だからといって、両者の意味は同じではない。むしろ二つの統計学の考え方には大きい隔たりがある。 P 値とは二つの治療の効果に差があるかどうかを測る物差しともいえる。但し厄介なことには P 値とは差がない確率そのものを素直に表してはいない^{3,16}。これに対して我々は「治療効果に差がないことはあり得ず、常に多少とも差がある」と考える。但し、一方の治療が常に他方より有効とはいえず、その逆が起こり得る。 P' 値は逆効果の起きる確率そのものを表す。良いと思ってやった治療が裏目にでることは、極力避けたいことであり、患者にとっても強い関心事である。 P' 値は P 値に比べてそれだけわかりやすい物差しである。これをもう一つの物差し、即ち「平均的治療効果の差」と組み合わせることにより、治療の選択がより確かなも

のとなる。

また二つの統計学の間には仮説そのものにも質的な違いがある。帰無仮説は否定されてもよい仮説であるが、第三の仮説は最も可能性が高く、推測の前提となる仮説である。両者では必要症例数の算定法も異なる。従来は二つのエラーの有意水準（ α 、 β ）をそれぞれどれだけを設定するかを決める必要があり、その根拠も明確ではなかった。これに対してエラーの有意水準（ α ）を一つ設定するだけで必要症例数は算出できる^{2,14}。またこのほうが容易であり、その意味も理解しやすい。さらにいえば、必要症例数の意味にも違いが生じる。従来は「有意な差を検出するために必要なサンプルサイズ」を意味したが、新しいコンセプトでは「エラーを冒すリスクが一定範囲に収まる集団の大きさ」となり、 P 値とグループサイズとは切り離すことができない。なおこの症例数の算定法は選択問題とほぼ同じである。

ちなみに1980年後半には医学統計学にある重要な変化がみられた。それは「 P 値や仮説検定をするよりは、信頼区間の推定を行うべきである」という勧告が医学のトップジャーナルを飾ったことである¹⁵⁻¹⁹。その中には「 P 値の終焉？」という題名で、「信頼区間には P 値の情報をすべて含むので P 値は不要」という論説さえみられた¹⁹。確かに臨床においては「差があるか否か」だけでは十分でなく、「その差がどの程度か」、更にいえば「その差が臨床的に意味のあるものか」を知ることは不可欠であり、この意味では P 値よりも信頼区間の推定のほうが重要である。

しかし、これだけで十分だろうか。母平

均の信頼区間が重視される他の理由は、これを「治療効果が同等か否か」という問題の決め手に使っているためであろう。例えば帰無仮説で「本当は二つの治療の効果が差がないが...」というときの「本当は」とは、「各治療群のパラメータ、特に母平均を集団同士で比較すれば」という意味である。しかし個人単位で2治療を比較すれば、たとえ全てのパラメータ同士が同じでも、各患者にとっては優劣がある（その差が無視できる程度のこともあるが）。患者や臨床医が知りたいのはこの個人にとっての優劣であるが、そこまではわからない。せめて優劣の割合だけでも推定することが望ましい。これが個人単位での P 値である。

従来の統計学では2治療の比較といえば、専ら集団同士の比較であり、その物差しとして集団を代表するパラメータが使われた。これを戦いに譬えれば、「勝敗は軍と軍の間の勝ち負けに限り、軍を代表する将軍をみて勝敗を判断すればよく、個々の兵卒にとっての勝敗までは考慮しない」という考え方と思われる。或いは兵卒の勝敗と将軍の勝敗とは混同されたか、少なくとも明確に峻別されなかったのではないだろうか。くどいようだが保険を例にしてみよう。説明すれば、従来の統計学は保険会社の視点でものを見たが、個々の被保険者の立場を考えなかったといえる。もっとも集団と個人とで利害の評価に違いがあることに気づいていた研究者や臨床医は稀とはいえない。例えば大規模臨床試験の結果を個々の患者の治療に当てはめることに注意を喚起するものはいたが、その違いを客観的に表現するまでには至らなかった。こうした問題に取り組むのが臨床統計学の課題であり、

P' 値は客観的な一つの尺度となる。しかしわれわれの臨床統計学はまだ完成した体系には至っていない。また目下のところ個人単位での P' 値は限られた条件下で、新たな仮説を立てなければ推定できない。

比率の群間比較

ここで新しい臨床統計学に基づき 2×2 分割表の検定を試みる。従来からこの検定には連続性の補正や Fisher の精密法その他の方法が発表されている。それぞれの方法によって P 値も異なる値をとり、どの方法を使うべきか利用者を惑わせることが多かった。

表 1 の例題は Armitage の教科書²⁰ から引用したもので、原典は Yates の論文²¹ である。母乳と哺乳瓶とで育てた幼児で咬合異常の発生率が $4/20$ と $1/22$ であった。この間に差があるだろうか。片側 P 値でみると、正規近似では $P=0.06$ 、Fisher の精密法では $P=0.14$ と大きい開きがある。

1. 集団にとっての P' の計算法

さて、平均すれば母乳で咬合異常が起きやすいが、その逆の場合もある。逆の結果が起きる割合が P' である。

今咬合異常をきたす幼児が、母乳では N_1 ($=20$) 人中 n_1 人 ($0 \leq n_1 \leq N_1$)、哺乳瓶では N_2 ($=22$) 人中 n_2 ($0 \leq n_2 \leq N_2$) 人である同時確率は、両者が独立に起きると仮定すると、

$$p(N_1, n_1; N_2, n_2) = \binom{N_1}{n_1} p_1^{n_1} (1-p_1)^{N_1-n_1} \binom{N_2}{n_2} p_2^{n_2} (1-p_2)^{N_2-n_2}$$

但し p_1, p_2 はそれぞれ母乳と哺乳瓶での咬合異常発生の母比率であり通常は、

$$p_1 = \frac{4}{20}, \quad p_2 = \frac{1}{22}$$

ただこの推定では例数が少ない場合に、発生率が 0 か 1 という両極端の値となると、 P' 値が不当に小さくなるおそれがある。これを防ぐためには以下のように事前確率が一様分布をすると仮定して、Bayes の定理から事後確率を求める。

$$p_1 = \frac{4+1}{20+2}, \quad p_2 = \frac{1+1}{22+2}$$

$P_1 > P_2$ の場合、 P' 値は例数の大小に拘わらず、以下の式で求められる。

$$P' = \sum_{\substack{n_1 < n_2 \\ N_1 < N_2}} p(N_1, n_1; N_2, n_2)$$

この値は 0.08 となり、Fisher の精密法よりは正規近似で求めた P 値に近い。

図 2 は以上の関係を立体ヒストグラムで図示したものである。図 1 のヒストグラムを二分するのが critical point であるのに対して、立体ヒストグラムを二分するものは critical line である。これは咬合異常発生率が母乳より哺乳瓶で多い領域（縞模様）と逆の領域（空白部）の境界線である。前者の領域内でヒストグラムの体積を加えたものと全体の体積の比が P' に相当する。但し本例では咬合異常発生率が母乳と哺乳瓶とで等しい場合が 3 箇所あり（打点部で、それぞれ 0, 0.5, 1.0）、この部分の体積は二分して、各領域に配分した。

注意すべきことは、比率の比較では第三の仮説から導いた P' 値と帰無仮説から求め

表 1. 幼児の咬合異常と哺乳との関係 (Yates)

	咬合異常	正常	合計	発生率(%)
母乳	4	16	20	20.0
哺乳瓶	1	21	22	4.5
合計	5	37	42	11.9

た片側 P 値とは同一ではない。

2. 個人にとっての P' 値

仮に上記の P' 値がもっと小さい値だとしても,それを基準にして母親は子供の咬合異常を予防するために,母乳を止めて哺乳瓶に代えるべきとはいえない.その理由の一つは, P' 値が20例と22例という集団から導いた値であるためである.敢えて一人一人にとっての P' 値を推定しようとするれば,厳密にはMcNemar検定のように,一人一人の乳児で母乳と哺乳瓶の両方を試みる必要がある.これは通常不可能である.但し,上述の方法と同様に咬合異常は母乳と哺乳瓶とで独立に発生すると仮定すれば推定は可能である.即ち,母乳で育てた乳児と哺乳瓶で育てた乳児とを総当り方式で比較して,咬合異常が前者で発生せず,後者で発生したペアの割合を計算する(但し咬合異常が両方で共に発生したペアと,共に発生しなかったペアの半分はこの中に加える).

$$P' = 0.5 - \frac{P_1 - P_2}{2} = 0.42$$

比率を見比べて治療を選択する場合には,期待値基準も P' も比率の差 ($p_1 - p_2$) に比例する.

3. 相対リスク(リスク比)の問題点

比率の差に対して,比率の比,即ち相対リスクが治療効果を評価する上で適当でないことを卑近な例で説明する.テロリストをかくまったという理由で,ある国が空爆を受けることになったとする.もし攻撃側が「今お尋ね者を引き渡せば空爆の範囲を半分に減らす」と提案したとすれば,守備側はその提案を無条件で受け入れて引渡しに同意するだろうか.これは「10万円を払えば新治療により死亡率が半分に減る」,つまり「相対リスクが1/2になる」というのに等しい.もし守備側の最高指揮者が愚かでなければ,その交渉に応じる前に確かめたいのは次のことであろう.即ち,もしお尋

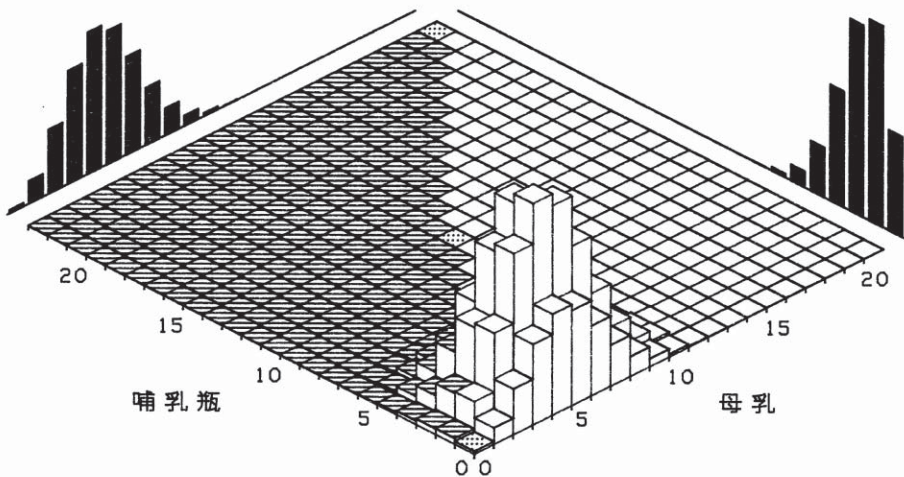


図2. 母乳20例と哺乳瓶22例の幼児における咬合異常例の同時発生確率
ヒストグラム上面の縞模様,打点,及び空白はそれぞれ哺乳瓶の咬合異常発生率が母乳より多いか,等しいか,少ない領域を表す.

表2. 旧治療に比べて新治療で障害発生リスクが半減したことの意味

全体的リスク	治療リスク		リスクの比	リスクの差	NNT*
	旧治療	新治療			
低場合	1/10 (10%)	1/20 (5%)	1/2	1/20	20
高場合	1 (100%)	1/2 (50%)	1/2	1/2	2

*Number needed to treat

ね者を引き渡さなかった場合はどれだけの範囲を空爆するつもりだったかである。もし最初の空爆予定が全領土の10%であれば、お尋ね者を引き渡せばその半分、つまり5%が破壊を免れる。ところがもし最初の空爆予定が全領土に及んでいれば、提案をのむことにより50%の領土が救われることになる。最初の空爆予定の範囲によって、破壊を免れる範囲が大きく変わる。つまり守備側は提案を拒否した場合と受諾した場合の空爆範囲の「比」だけでは納得せず、「差」を知ろうとするのが普通である。

同様に治療効果の違いを死亡率の比で表しても損得の程度は把握できない。本来はリスクの差で表すべきである。つまりリスクの比でいえば同じ1/2でも、差になおせば、死亡リスクは5%のこともあれば50%に達することもあるのである。これは新治療で余分に助かる患者の割合であり、それぞれ1/20と1/2に相当する(表2)。

4. NNTの提唱

ところが上に述べた単位のない分数ではまだ損得が理解しにくいといわれる。もっと損得がわかりやすいのは、金銭、年月、人数などの単位のついた数値で示すことである。例えば上の例で、「新治療で余分に助かるのは20人中の1人であり、残りの19人はどちらの治療をしても結果は同じである」

といったほうが、10万円を払って新治療を受けるべきかどうかの判断がしやすい。言い換えれば1/20ではなく、その逆数を使う方法である。この指標は1985年に直腸癌根治手術における下腸間膜根部のリンパ節廓清の得失を検討するために使用された²²。その結果、このリンパ節の廓清で余分に助かるのは500人中僅か1人に過ぎなかった。1988年これがNNT(Number needed to treat)という名の下にLaupacis, Sackettらによって提唱され、一躍注目を浴びるに至った²³。

打切りのある生存時間の群間比較

1. 相対ハザード(ハザード比)の問題点

これは現在もRCTにおける生存分析で屢々使用され、治療の有効性を統計学的に示す根拠となっている。ちなみにlogrank検定結果はハザード比の区間推定と一致する。ハザード比は相対リスクよりももっと誤解を招きやすい指標である。これを再び空爆を例に説明する。

「お尋ね者を引き渡せば空爆の範囲を半分に減らす」という攻撃側の提案を守備側が受諾して、お尋ね者を引き渡したとする。ところが攻撃側はそれでも毎日空爆を続け、国全体が焦土となった。守備側はこれは明らかに約束違反であると抗議した。すると、攻撃側の言い分は次の通りである。「空爆を

一日で終結するという約束をした覚えはない。もし提案を拒否した場合は、まだ破壊されていない残っている領土(図3の白い部分)の1/2を毎日破壊する予定であった。提案を受諾したので、破壊する範囲(黒い部分)を1/2からその半分の1/4に減らしている。どの日をもこの約束は遵守されているのではないかと。これに対して守備側は更に反論する。「たとえ空爆を毎日続けることを認めるとしても、その範囲を半分にするといえ、最終的に破壊の範囲が半分に減るものと解釈するのが普通である。遅かれ早かれ全領土を破壊するならば、その範囲を半分に減らすなどというのは詭弁にほかならない。お尋ね者を引き渡せば、破壊を何日間か先に延ばすというべきである」。実は相対ハザード(ハザード比)とはこの破壊が半分に減るといふ話と全く同じで、誤解を招きやすい。ハザードが半分になったからといって遅かれ早かれ全員が死亡

するかもしれないのである。にも拘わらず、ハザード比が1/2となったことを「死亡率が半減した」とか、「半数を救命した」のように記述する論文は少なくない。BMJが発行するEBMの二次情報誌Clinical Evidence 7(2002)の始めに統計用語の解説がある。その中の「ハザード比」を見ると、「広義の相対リスク(RR)である... (中略)...もしこれが0.5ならば、一方の群で死亡する相対リスクは他の群で死亡するリスクの半分である」と書かれている。これでは単に死期が後にずれたのを死亡が半減したものと誤解しやすい。Tanは延命と救命とを混同することを戒め、「将来は治療の効果を発表する論文に対して、編集者はどれだけ生存期間を延ばしたかを記載するよう要求すべきである」と提案している。²⁴ 図3では空白の部分の面積が平均生存時間に相当し、提案を受諾することによりこれが延長している。

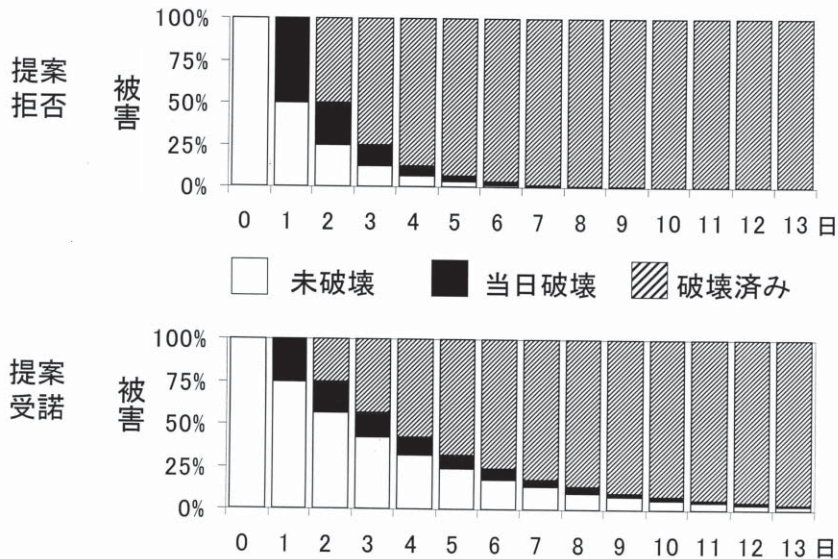


図3. 提案を拒否した場合と受諾した場合の空爆による破壊の程度の違い
 上下段のヒストグラムの空白の部は二つの生存曲線に相当し、1日当たりのハザードは上段の1/2から下段の1/4に減少している(ハザード比1/2)。

2. 治療効果を評価するための平均余命の意義

かつては平均生存時間（平均余命）を推奨する統計学者は極めて少なかった。例えば1977年 British Journal of Cancer で Peto, Armitage, Breslow, Cox, Mantel など世界的に有名な生物統計学者 10 名が連名で RCT の方法を平易に解説している。²⁵ この中で彼らが最も推奨したのは logrank 検定である。一方 50% 生存時間や 5 年生存率には問題があるが、平均生存時間はそれ以上に悪い指標であり、使用すべきでないとして述べている。これはわれわれの評価と全く逆である。^{3,5,26} しかし 20 世紀の終わり頃から平均余命を支持する研究者は次第に増加している。^{24,27-30}

ただ生存期間の平均値よりも中央値を推奨する統計学者は多い。その根拠は生存期間が右側に長い裾を引く非対称性の分布をするためである。^{31,32} このような分布を要約する場合には、平均値 ± 標準偏差（または標準誤差）は不適切であり、箱ひげ図やパーセント点を使うべきといわれている。とすれば 50% 点である中央値は分布の要となる。にも拘わらず敢えて平均値を推奨するにはいくつかの理由がある。²⁵⁻²⁹ われわれの理由の一つは既に述べたように、集団の利害しか評価できないのであれば、平均値が最適の指標であるからである。もう一つの理由は中央値や 5 年生存率はいずれも生存曲線上の一点の x 値または y 値であるが、この点は critical point とは言いがたく、むしろ恣意的に選んだ点である。一方、平均値は生存曲線下の面積（AUC）であり、理論的にもより望ましい指標といわれる。³⁰

図3において白い空白の部分から右に辿っていくと Kaplan-Meier の生存曲線が描ける。つまり生存曲線とは生存率や死亡

率を表す縦棒の集まりとみなされる。ところが図4に示すように生存曲線には別の見方がある。即ち生存時間を表す横棒を長いものから短いものの順に積み重ねたものが生存曲線となる。これから AUC が平均生存時間であることは容易に理解できる。

3. 数理モデルを用いた平均余命の推定

平均余命の最大の難点は、これを実測しようとするれば、最長生存患者が死亡するまで待たねばならないが、その前に研究者のほうが先に死亡するおそれがある。短期追跡データから平均余命を推定するためには、数理モデルを用いて、将来の生存曲線を予測しなければならない。これまで数多くのモデルが提唱されたが、われわれの知る限り、あらゆる生存曲線に適合するような柔軟なモデルはない。その理由は、生存曲線の形を変える因子は原疾患の進行度や悪性度だけではないからである。原疾患による早期死亡を免れた患者の長期生存曲線は、

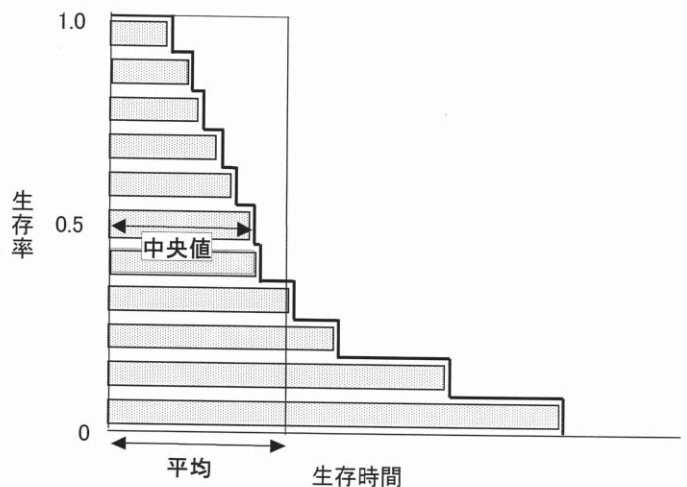


図4. Kaplan-Meier の生存曲線の別の見方
平均生存時間は生存時間を表す横棒の和となり、
曲線下の面積と一致する。これは同じ面積の長
方形の横の長さ に等しい。

患者の年齢,性,年代,地域などによって,大きい影響を受けるが,後者をモデル化することは難しい.そこでわれわれは生存曲線を原疾患に關係する成分と,これと無關係な成分に分け,前者にはBoagのモデル³³を適用し,後者は患者と同年代,同年齢,同性の日本人の世代生命表で近似させた.こうして互いに独立とみなされる両成分を再合成することにより,生存曲線や平均余命を推定した.^{5,26,34}本モデルを本院腹部一般外科³⁴及び癌研外科³⁵で手術した胃癌患者の長期追跡例に適用した結果,5年追跡の時点でそれ以後の曲線がCoxモデルよりも遙かに正確に予測でき,また平均余命の予測も正確であった.更にBoagモデルを重回帰モデルにまで拡張することにより,与えられた群の平均余命ではなく,与えられた予後因子をもつ患者の平均余命を求めることも可能となっている.^{5,35}

費用対効果比

こうして平均余命が推定可能となると,治療群間での平均余命の差が求められる.これはADLG (average duration of life gained)と呼ばれ,5年生存率の差やハザード比に代わって治療のアウトカムを評価するための新しい指標となり得る.³⁶しかし医療費高騰の現在では,新しい医療技術で寿命がどれだけ延長したかだけでは,その技術を採用するのに十分とはいえない. Relmanは米国における医療費抑制の時代(the era of cost containment)の後に来るべき時代として,評価と説明責任の時代(the era of assessment and accountability)を挙げている.³⁷この時代においては医療提供者は新しい医療技術によって1年寿命を延ばすためには

どれだけ余分の費用を要するかを提示し,支払い者側を納得させる必要がある.このための一つの指数が費用対効果比(cost-effectiveness ratio)である.³⁸⁻⁴⁰米国では新技術採用の条件として,この比が5万ドル/年以下,英国では3万2千ドル以下ともいわれる.²⁹多くの費用と労力及び被験者を要するRCTもこの評価の対象となる.⁴¹

おわりに

今や臨床統計学に求められているのは,ある医療技術が他の技術と比べて有効性に差があるというエビデンスを示すことだけでは十分ではない.その医療技術が費用を含めて導入に値するかどうかを消費者のわかる言葉と物差しを使って正確に説明し,納得させなければならない.これは極めて難しいが,その原因はどこにあるのだろうか.アメリカのある病院の手術場に「もし手術が難しいと感じたならば,それはあなたが間違っただけをされている」と書かれていたという.同様に,説明が難しいのは正しいことをやっていないためではないだろうか.この意味では臨床統計学が志向するのは説明のできる医療,即ち evidence-based medicine を超えて accountable medicine というべきかもしれない.

謝辞

貴重なコメントを頂いた武田薬品工業統計解析部の森川敏彦部長と東京大学大学院薬学研究科医薬経済学の津谷喜一郎教授に深謝する.

参考文献

1. Fahey T, Griffiths, Peters TJ. Evidence-based purchasing: understanding results of clinical trials and systematic reviews. *Br Med J* 1995;311: 1056-1060.
2. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *J Chron Dis* 1967;20:637-648.
3. 天理よろづ相談所医学統計解析グループ . 情報の価値からみた医学統計学の再検討 . 臨床評価 1999;27:393-407.
4. 天理よろづ相談所医学統計解析グループ . 証拠に基づく医療 (Evidence-based medicine): その妥当性と限界 . 天理医学紀要 2000;3: 138-153.
5. 天理よろづ相談所医学統計解析グループ : 臨床側からみた生存分析:Coxモデルから新しい生存モデルへ . 天理医学紀要 2001; 4: 128-146.
6. Bunn DW. Analysis of optimal decision. John Wiley & Sons, Chisester, 1982.
7. Weinstein WC, Fineberg HV. Clinical decision analysis. WB Saunders, 1980.
8. Kahneman D, Tversky A. Choices, values, and frames. In : Kahneman D, Tversky A. eds. Choices, values, and frames. New York: Cambridge University Press; 2001:1-16.
9. Sackett DL, Straus SE, Richardson WS, et al. Evidence-based medicine: How to practice and teach EBM. 2nd Ed. Churchill Livingstone, London, 2000.
10. Sackett DL, Rosenberg WMC, Gray JAM, et al. Evidence-based medicine: what is it and what it is not: It's about integrating individual clinical expertise and the best external evidence. *Br Med J* 1996;312:70-72.
11. Maetani S. Another approach to clinical trial numbers. *Lancet* 1990;335:114.
12. Nichol JP. Clinical trial numbers. *Lancet* 1990; 335:483
13. Fontbonne A, Schwartz D. Clinical trial numbers. *Lancet* 1990;335:483.
14. 前谷俊三 .臨床生存分析: 生存データと予後因子の解析 . 南江堂 , 東京 1996:81-100.
15. Simon R. Confidence intervals for reporting results of clinical trials. *Ann Intern Med* 1986;105: 429-435
16. Ware JH, Mosteller F, Infelfinger JA. P values. In: Bailar JC, Mosteller F. eds. Medical uses of statistics. Boston, the Massachusetts Medical Society; 1986. 津谷喜一郎 , 折笠秀樹 監訳 P値 .医学統計学の活用 .東京 :サイエンティスト社 ; 1995:125-141.
17. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986;292:746-750.
18. Bulpit CJ. Confidence intervals. *Lancet* 1987; 494-497.
19. Evans SJ, Mills P, Dawson J. The end of p value? *Br Heart J* 1988;60:177-180.
20. Armitage P. Statistical inference. In: Statistical methods in medical research. Oxford: Blackwell Scientific Publications; 1971:99-146.
21. Yates f. Contingency tables involving small numbers and the χ^2 test. *J Roy Statist Soc Suppl* 1: 217-235.
22. 前谷俊三 , 飯島庸介 , 戸部隆吉 . 直腸癌における拡大根治手術とその評価 . 外科 1985; 47:147-152.
23. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measure of the consequences of treatment. *New Engl J Med* 1988; 318: 1728-1733.
24. Tan LB, Murphy R. Shifts in mortality curves: saving or extending lives? *Lancet* 1999;354:1378-1381.
25. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer* 1976;34:585-612.
26. Maetani S, Yamazoe Y, Onodera H, Imamura M. Estimation of mean survival time from limited follow-up in cancer patients. *Tenri Medical Bulletin* 1998;1:38-49.
27. Wright JC, Weinstein MC. Gains in life expectancy from medical interventions: standardizing

- data on outcomes. *New Engl J Med* 1998;339:280-286.
28. Messori A, Becagli P, Tripoli S. Median versus mean lifetime survival in the analysis of survival data. *Oncol Report* 1999;6:1135-1141.
 29. Beard SM, Holmes M, Price C, et al. Hepatic resection for colorectal liver metastases: a cost-effective analysis. *Ann Surg* 2000;232:763-776.
 30. Lau EW. Visual illusions created by survival curves and the need to avoid potential misinterpretation. *Med Decis Making* 2002;22:238-244.
 31. Altman GA, Gore SM, Gardner MJ, et al. Statistical guidelines for contributors to medical journals. *Br Med J* 1983;286:1489-1493.
 32. Murray GD. The task of a statistical referee. *Br J Surg* 1988;75:664-667.
 33. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J Roy Statist Soc B* 11:15-53, 1949
 34. Maetani S, Matsusue S. Mean survival time in gastric cancer patients estimated from a parametric model: comparison with long-term follow-up results. *Tenri Medical Bulletin* 2000;3:15-25.
 35. Maetani S, Nakajima T. 30- to 50-year follow-up of gastric cancer patients. Are the results predictable five years after surgery? *Proceedings of the 4-th International Gastric Cancer Congress*, Munduzzi Editore, Bologna, 2001.
 36. Lubsen J, Hoes A, Grobbee D. Implications of trial results: the potentially misleading notions of number needed to treat and average duration of life gained. *Lancet* 2000;1757-1756.
 37. Relman AS. Assessment and accountability: the third revolution of medical care. *New Engl J Med* 1988; 319:1220-1222.
 38. Russeli LB, Gold , Siegel JE, et al. The role of cost-effective analysis in health and medicine. *JAMA* 1996;276:1172-1177.
 39. Weinstein MC, Siegel JE, Gold MR, et al. Recommendations of the panel on cost-effectiveness in health and medicine. *JAMA* 1996;276:1253-1258.
 40. Siegel JE, Weinstein MC, Russel LB, et al. Recommendations for reporting cost-effective analyses. *JAMA* 1996;276:1339-1341.
 41. Greenfield S. The state of outcome research: are we on target. *New Engl J Med* 1989;320:1142-1143.

From Neyman-Pearson statistics to a new clinical statistics: accountability rather than evidence

Statistical Analysis Group

Tenri Institute of Medical Research

In the present climate of increasing demands for patient-centered health care, for informed choice of treatments and for professional accountability, medical statistics needs change. Here, we indicate some shortcomings of Neyman-Pearson statistics and propose a new clinical statistics which is more comprehensible to patients and society, helping them to make better decisions. The primary criterion for selecting a better treatment is the mean difference in treatment efficacy (expected value criterion). However, the treatment selected has a risk of producing a worse outcome than the alternative treatment, so that we also need to estimate this probability (P'). It is approximately equal to one-sided P value, assuming that the sample mean and variance represent the population parameters (the third and most likely hypothesis). This risk is known as the γ error (Schwartz), and is the patient's primary concern, in contrast to the α and the β errors. The P' value for individual patients should be distinguished from that for the group; clinical decision making is refined by combining the expected value criterion with the former P' value. To evaluate the treatment efficacy in survival time analysis it is better to parametrically estimate the mean survival time difference rather than the hazard ratio, so that cost-effectiveness ratio is also derived.

Keywords: Evidence-based medicine, professional accountability, expected value criterion, γ error, mean survival time