

臨床試験における多重性の問題について

アステラス製薬株式会社 データサイエンス部 柿爪 智行(Tomoyuki Kakizume)
Data Science Dep. Astellas Pharma Inc.

1. はじめに

臨床試験においてしばしば問題となるものの1つに、多重性(multiplicity)が挙げられる。これは、複数の検定を同時に行うことにより、試験全体としての第I種の過誤(type I エラー)が望ましい水準よりも増大してしまう現象である。これを解決するために、多重比較法が考えられてきた。これは、「試験全体として」の type I エラーを望ましい水準にコントロールするために、各検定における棄却限界値を調整する手法である。

現在はさまざまな多重比較法が存在しているが、例えば副次変数を複数設定する試験や部分集団解析等、用いる解析手法や結果によっては解釈が困難となる場合がある。このようなとき、どの手法を用いればよいか、またどう解釈すべきかといった問題に関しては、明確と思われるルールが必ずしも存在せず、日・米・EU 三極医薬品規制調和国際会議(ICH)から出されたガイドライン「臨床試験のための統計的原則」(ICH-E9) [2]においても、このような観点からは言及されていない。

一方、欧州医薬品審査庁 EMEA (European Medicines Evaluation Agency. 2004 年 5 月に European Medicines Agency に改名)の欧州医薬品委員会 CPMP (Committee for Proprietary Medical Products. 2004 年 5 月に Committee for Medical Products for Human Use に改名)から複数の Points to Consider が公表されている。これらの位置付けは、ICH-E9 で十分に議論されていない事項についての手引きであると考えられる。本発表では、その中で 2002 年 8 月に公表された POINTS TO CONSIDER ON MULTIPLICITY ISSUES IN CLINICAL TRIALS (以下 PtC と略す) [1] を取り上げ、その内容を紹介するとともに、臨床試験における多重性の問題についての論点及び意見を示す。

2. PtC の内容紹介

2.1. INTRODUCTION

PtC では、通常起こる多重性に対する懸案事項として「多重性が適切に処理されなかった場合、薬の有効性に対する確証のない主張に対して、正しいのに誤っているという結論を導いてしまう確率(type I エラー)が増加してしまう」ことを挙げている。よって、特に検証的な臨床試験の評価において重要なことは「試験全体の type I エラーを容認できる水準 α にコントロールすること」であるとしている。ここでの「type I エラー」とは、「複数の帰無仮説のうち、少なくとも1つが正しいにも関わらず誤って棄却されてしまう確率」を指す。

PtC では、最近の欧州での申請における多重性に関する重要な問題として、以下について主に

解説されている。

- ・ 多重性の調整の必要性
- ・ 副変数から得られた「有意」の解釈およびこれらに基づいた主張
- ・ 部分集団解析から得られる結論および特定の部分集団に対する承認の制限
- ・ 「レスポnder」を用いた解析の解釈
- ・ 複合変数の定義・解析・解釈

以下では、上記の点について順に紹介する。

2.2 多重性の調整の必要性

type I エラー α を調整するとは、(いつもではないが) しばしば予め定められた α を分割し、それらを実際に行う複数の帰無仮説検定の水準にそれぞれ設定することを意味する。 α を調整する際に注意すべき点として、以下が挙げられている。

- ・ 多重検定を用いると、臨床的な解釈が難しくなる。そのため、用いる手法を選択する際には、十分に臨床的な解釈を考慮に入れた妥当な統計的手法が存在するかどうか考えること。
- ・ 多重性の調整法にはさまざまな種類があり、用いた手法が α を正しく調整するものであっても、その手法によっては異なる結論を導くことがある。そのため、治験実施計画書(プロトコル)または統計解析計画書に用いる手法を予め記載しておくこと。
- ・ 多重性の調整が必要かどうか予見できず、事後的に用いた場合には保守的な方法(Bonferroni 法等)を用いること。

2.2.1 多重性の調整を必要としないケース

複数の主要変数を設定した検証的試験においても、 α の調整を必要としないケースが存在する。それらを以下に述べる。

- ・ 複数の主要変数を用いて臨床的な治療効果を検証する必要があるケース

薬剤によっては、複数の主要変数を設定し、それらに対して臨床的な治療効果を検証する試験を行うことがある。その際、設定された変数全てに対して統計的有意差を示すことが必要となるが、この場合 α の調整は必要でない。ただし、関連するタイプ II エラー(ここでは、誤った帰無仮説を少なくとも1つ採択してしまう確率を指す。以後、 β と表す)は増加してしまうため、試験に必要な症例数を見積もる際にはこのことを考慮に入れなければならない。

- ・ 臨床的な関連性に従って複数の主要変数が順序付けされるケース

主となるべき結論に関連した複数の副次的な結論が認められたとき、治療の価値が上がることもある。典型的な例としては、(i) 抑うつ障害の急性効果に関連した障害の進行の防止、(ii) 急性心筋梗塞による死亡の減少に関連した他の重篤なイベントの防止などがあげられる。その際、各変数の重要度または治験責任医師の関心度の大きさなどによってそれぞれに順序付けを行い、それに従って段階的に仮説検定を行うことがある。この場合 α の調整は必要ないが、順序付けの方法に関しては試験実施計画書に予め記載しておかなければならない。

この場合、棄却されなかった仮説およびそれより低い順序の仮説については、検定を行った結果統計的有意差が認められたとしても、検証されたことにはならない。また、より低い順序の仮説について β は増加することに注意すべきである。

2.2.2 安全性項目における多重性

安全性の検証が試験目的の1つだった場合、観測された結果により、薬物の悪影響が示されたり懸案事項が生じたりする場合を除いては、有効性に対する主要項目と同様の扱いをする。ただし、副作用が起こった場合、P値の大きさ(発現率の差)よりも、発現した副作用の重症度・重篤度や転帰に依存した、実質的なリスクの大きさが重要となる。

安全性の評価をする際、治験薬の潜在的なリスクに対して多くの検定を用いることがある。このとき α の調整を行うと、各検定に用いる水準が小さくなるため、帰無仮説が棄却されにくくなり、結果的に治験薬の安全性を行う上で非常に重要視されるべき「投与群間で差がある」という結論を導きにくくなる。以上の点から、安全性の評価をするために複数の検定を行う場合には、 α の調整を行うべきではないと考えられる。

2.2.3 多群の比較における多重性

複数の治療群を設定して検証試験を行う際、一般的に α の調整が必須であるが例外も存在する。ここでは、3つのデザインを例に挙げて、それぞれの場合において多重性の問題の扱い方を述べる。

2.2.3.1 3群「gold standard」デザイン

標準薬が存在する疾患の場合、3つの治療群(プラセボ、標準薬、治験薬)を用いて治験薬の有効性および安全性を示すことが薦められている。このような試験の目的は、以下の3点である。

- (1) プラセボに対して、治験薬の優越性を示す(有効性の証明)
- (2) プラセボに対して、標準薬の優越性を示す(感度分析の証明)
- (3) 標準薬に対して、治験薬が劣っていないことを示す(非劣性の証明)

このデザインを用いる際、(1)~(3)全てにおいて統計的有意差を示さなければならず、このとき α の調整は必要でない。

プラセボに対して治験薬の優越性を示せなかった場合、「治験薬が有効でない」(プラセボに対して標準薬の優越性が示せた場合)、または「感度分析の欠如」(プラセボに対して標準薬の有効性が示せなかった場合)と解釈される。

2.2.3.2 併用試験における有効性の証明

CPMPガイドラインによると、例えば2種類の治療を併用することにおける有効性を証明するためには、2種類の治療を併用する群と各治療を単独で行う群の計3群を用いて、各単独使用群に対してそれぞれ併用使用群の優越性を示すことが必要となる。この場合、 α の調整は必要ない。

2.2.3.3 用量反応試験

治験薬の推奨用量の探索・確認のために、複数の用量群を設定して用量反応試験を行うことがある。その際、 α の調整は必須となる。用量反応試験における仮説や目的(例：単調性の仮定の有無、最小有効用量または標準薬の推奨用量に対して同様な(劣っていない)用量の探索)について、さまざまな手法が文献で紹介されているが、PtCではそれらに関してはとくに言及されていない。

検出力不足のため有効かつ安全であると考えられる推奨用量(あるいは推奨範囲)を特定できなくても、用量の増加に対して臨床的な効果が正の相関を示すことが認められれば、その結果は有用であると考えることができる(ICH-E4[3]参照)。

次試験に用いる用量を決定するために、得られたデータを用いて予備解析を行う場合がある。その際、対比較を用いた推定や信頼区間の算出が行われるが、 α の調整は必要ない。

ICH-E9では、臨床試験において「主要変数は通常ただ1つにすべきである」としている。ただし、1つの変数では臨床的に適切な治療効果を十分観察できない場合には、複数の主要変数を設定することが必要になる。この場合、各変数の対立仮説や検出力限界の定義は、お互いバランスを保たなければならないため、必要症例数の算出はさらに複雑になる。

このような事態に対する対処法としては、事前に治療効果や対象疾患の特徴を調査した上で、解析に用いる変数を必要最小限に抑えることが考えられる。また、今後の課題として、上記の場合に適応する理論的な症例数の算出方法の導出が挙げられる。

2.3 副次変数から得られた「有意」の解釈およびそれらに基づいた主張について

2.3.1 補助的な証拠を得るための変数

主要変数を補助する目的として設定された副次変数はあくまで予備的なものであるため、そこから得られた信頼区間および統計的検定の結果もまた位置付けとしては予備的なものとなる。

2.3.2 追加の主張の基礎となりうる副次変数

副次変数は、「補助」としての役割の他に、主要変数に対して「追加の」主張を導く役割ももつ。この場合、副次変数は主要変数より下の階層として扱う。また、副次変数による主張は、まず主要変数による主張が認められた場合にのみ、主要変数の主張に対する追加の主張として認められる。

2.3.3 臨床効果を示す変数

死亡など、重要な安全性の問題と考えられる事象を変数に指定する場合、変数の特性上計画される試験の検出力が小さすぎる恐れがあるため、副次項目として設定することがある。この場合、主要変数において統計的に有意な効果が認められたとしても、本来は主であるこの項目において統計的に有意な効果が認められなければ、追加の試験が必要になる場合がある。

2.4 部分集団解析から得られる結論および特定の部分集団に対する承認の制限

性別、年齢、生活地域、対象疾患の重篤度、人種、薬物の吸収や代謝機能の違いなど、さまざまな要因が治療効果の不均質を起こすことが知られている。治療効果に影響を及ぼすという疑念が事前にはない限りは、通常これらを層とした部分集団解析は探索的または予備的なものと考えられる。

ある特定の部分集団における治療の効果を明確に述べるためには、適切な解析計画を予め定めておくことが必要である。その際、治験実施計画書で検出力の検討を行うことや、層ごとに無作為化を行うことが望ましい。

ただし以下の場合、特定の部分集団において追加の臨床データが得られるまで承認を制限されることがある。

- ・ 治療に悪影響を及ぼす強い相互作用が示され、この現象に対して説得力のある説明が得られないとき、また他からの情報によりその相互作用が確認されたとき
- ・ 歴史的に、ある部分集団の患者は薬の効果が得られないと信じられており、得られた結果がその考えと強く矛盾しないとき
- ・ 解析計画書において適切な解析計画を記載せずに、さまざまな部分集団解析を行ったとき
- ・ 試験で用いる母集団の重要な定義が欠如しているとき

2.5 「レスポナー」を用いた解析の解釈

アルツハイマーやてんかんなどの疾患において、平均すると小さいが統計的に有意となるような改善効果を解釈するのは困難である。このような場合に、各患者に対する治療の臨床的效果を評価するためレスポナー/ノンレスポナーという概念が用いられる。

【レスポナー/ノンレスポナーの概念】

連続型または離散型確率変数 X に対してある条件 A を設定し、次のような関数 f を定義する。

$$f : \text{連続値 or 離散値} \mapsto [0, 1],$$

$$f(X) = 1(X \in A); 0(X \notin A)$$

このとき、

$$X \text{ がレスポナー} \quad \Leftrightarrow f(X) = 1$$

$$X \text{ がノンレスポナー} \quad \Leftrightarrow f(X) = 0$$

レスポナー/ノンレスポナーの条件 A の設定方法は試験によって異なるため、どの方法を用いたかということについて治験実施計画書内で予め定めておくべきである。また、用いる設定方法は臨床的な意味も持つべきである。ただし、レスポナー/ノンレスポナーを用いる際には、変数の大きさや変数間の差など、元の変数に含まれていた情報の損失(それゆえ検出力の低下)が起こることに注意すべきである。

レスポナー/ノンレスポナーを用いた解析においては、統計的に有意かどうかは必要でなく、レスポナー率の差が評価すべき治療の適切な臨床的效果を裏付けることが必要である。なお、この解析は主要変数に効果が見られなかった場合にそれを救済するようなものではないことに注

意すべきだ。

個人的には、レスポonder・ノンレスポonderを用いるのは以下の場合がよいのではと考えている。

- ・ 変数間の「差」が効果の「差」と一致しない場合
以下の質問及び回答を例に考える。

〈質問〉

観察期と比べ、現在の症状はどの程度改善しましたか？

〈回答〉

0=症状がなくなった

1=かなり改善した

2=やや改善した

3=変わらなかった

4=悪くなった

この回答のように、数値の大小で順序関係を示す変数を順序尺度変数と言うが、この変数は数量的な情報をもたないため、各変数の「差」は何の意味ももたない。このような場合には、ある基準を設定してレスポonder/ノンレスポonderに変数変換を行うことにより、2値の解析に移行することが薦められる。

- ・ 変数間のわずかな「差」が臨床上重要な「差」である場合
わずかではあるが、それが臨床上重要な変化となるような項目を解析する際、その項目を連続値として解析すると統計的に有意な差が認められない場合がある。このような場合、ある基準を設定してレスポonder/ノンレスポonderに変数変換を行い2値の解析に移行することが薦められる。ただし、レスポonder/ノンレスポonderを用いた解析は救済措置として用いるべきではなく、基準に関しては臨床的な観点から設定すべきである。

2.6 複合変数の定義・解析・解釈

PtCでは、とくに生存解析に関する複合変数について解説されている。

通常、生存解析で扱う事象が起こることは比較的稀であり、それぞれを単独の事象として評価する場合非常に多くの症例数が必要になる。それゆえ、あらかじめリスト化された複数の事象(例：死亡、心筋梗塞、無能力状態)のうちどれかが発現するかどうか、また発現するまでの時間を評価項目に設定することがある。

2.6.1 主要項目としての複合変数

有効性を示すために複合変数を用いる場合、通常それが主要変数になる。そのため、承認の際必要となる重要な有効性を示す証拠をその複合変数が供給できなければならない。加えて、補助

的な情報を得るために、各事象について、また関連のある事象をまとめた解析をそれぞれ行うことが薦められる。

2.6.2 複合変数の定義

複合変数を定義する際、行う治療が影響すると考えられる要素のみを含むことが薦められる。また、非劣性試験において複合変数を用いた場合、含めた変数中のある要素で治療の悪影響が観測されたときには解釈が困難になることに注意すべきである。

2.6.3 複合変数の解析

複合変数を用いた解析から得られる結果に対する懸念事項として、以下の点が挙げられる。

- ・ 治療により 1 つまたは複数の事象に対して悪影響を及ぼされた場合、複合変数ではどれに起きたか特定できない。
- ・ 特定の事象に悪影響が起きたとしても、他の事象に対しては治療の効果得られた場合、その悪影響が隠れてしまう。

2.6.4 複合変数の解釈

複合変数を「死亡および肝移植の発現」としたが、実際には死亡がごく少数にしか発現しなかった場合を考える。設定した複合変数を用いて解析を行った結果、統計的に有意な効果が見られたとしても、複合変数に対して「死亡」から得られる情報は小さいと考えられるため、「死亡および肝移植の必要性の減少」という結論を導くことはできない。

2.7 結論

臨床試験において、新薬の開発が成功するか否かは肯定的な結論に依存するため、治療の有効性(安全性)に関する複数の問題に答えることがしばしば必要になる。ただし、一つの試験から多くの問題に答えようとする際、多重性の問題が発生する。多重性の問題を含んだ試験を扱う、または避けるために、適合性や妥当性を評価するために予め計画された統計的手法を治験実施計画書または統計解析計画書に十分詳しく述べる必要がある。

また、副変数の解析または部分集団解析から得られる統計的有意または臨床的に意味のある主張は、それらが予め特定された適切な統計解析計画に基づいて解析が行われており、主要変数による主張が認められた場合、追加の主張として捉えることができる。

3. 問題提起

ここでは、前章で紹介した PtC を踏まえた上で、個人的に感じた疑問点および問題点を挙げる。

3.1 多重比較における検出力および理論的な症例数の算出

複数の検定を同時に行う多重比較では複数の帰無仮説が存在するため、検出力の概念の理解も難解である。ここでは、実際の治験で問題になると考えられる 2 種類の多重比較(多群の比較、

複数の評価項目を設定した比較)それぞれにおける検出力について考察する。

3.1.1. 多群の比較における多重比較

ここでは、多群の比較を行う際に生じる多重比較における検出力について、「統計的多重比較法の基礎」[8]に従って考察する。

多重比較においてはさまざまな検出力の定義が存在する。例として、以下のものがあげられる。

- (a) 総体検出力：母平均間に差のある全ての対を検出する確率
 - (b) 制約付総体検出力：母平均間にある一定の値以上の差がある全ての対を検出する確率
 - (c) 最小差検出力：母平均間の差が最小である(ゼロの場合は除く)対を検出する確率
 - (d) 最大差検出力：母平均間の差が最大である対を検出する確率
 - (e) 平均棄却数比：母平均間に差のある全ての対のうち検出された対の割合
 - (f) 重み付き平均棄却数比：(e)において母平均間の差で重み付けをした割合
- また、これらを式で表すと、それぞれ以下ようになる。

第 i 群の母平均を μ_i ($i=1, \dots, a$)、第 i 群と第 j 群に対する帰無仮説を $H_{(i,j)}$:

$\mu_i = \mu_j$ ($1 \leq i < j \leq a$) で表す。このとき、(a) ~ (f) は次式で表される。

(a) 総体検出力：

$$\Pr\left\{\sum_{(i,j) \in I_1} (1 - D_{(i,j)}) = 0 \mid (\mu_1, \mu_2, \dots, \mu_a)\right\}$$

ただし、

$$D_{(i,j)} := \begin{cases} 1 & \text{(帰無仮説 } H_{(i,j)} \text{ が棄却された場合),} \\ 0 & \text{(帰無仮説 } H_{(i,j)} \text{ が採択された場合),} \end{cases}$$

$$I_1 := \{(i, j); 1 \leq i < j \leq a, \mu_i \neq \mu_j\}$$

(b) 制約付総体検出力：

$$\Pr\left\{\sum_{(i,j) \in I_2} (1 - D_{(i,j)}) = 0 \mid (\mu_1, \mu_2, \dots, \mu_a)\right\}$$

ただし、

$$I_2 := \{(i, j); 1 \leq i < j \leq a, |\mu_i - \mu_j| > \Delta\} \quad (\Delta : \text{ある一定の値})$$

(c) 最小差検出力：

最小差をもつ対を (i, j) としたとき、

$$\Pr\{D_{(i,j)} = 1 \mid (\mu_1, \mu_2, \dots, \mu_a)\}$$

(d) 最大差検出力：

最大差をもつ対を (i, j) としたとき、

$$\Pr\{D_{(i,j)} = 1 \mid (\mu_1, \mu_2, \dots, \mu_a)\}$$

(e) 平均棄却数比：

$$E\left\{ \sum_{(i,j) \in I_1} D_{(i,j)} \mid (\mu_1, \mu_2, \dots, \mu_a) \right\} / \#I_1$$

(f) 重み付き平均棄却数比：

$$E\left\{ \sum_{(i,j) \in I_1} D_{(i,j)} \mid (\mu_1, \mu_2, \dots, \mu_a) \right\} / \left\{ \#I_1 \sum_{(i,j) \in I_1} |\mu_i - \mu_j| \right\}$$

臨床試験を行う上で検出力が問題となるのは、試験の計画段階においてサンプルサイズを設計するときである。2群の比較を行う際、検出力は、「臨床的に意味のある差を設定し、比較する変数の差がそれを超えている場合に正しく「比較群間に差がある」という結論を出す(帰無仮説を棄却する)確率」を指すことはよく知られている。この観点から考えると、実務的には(b)が最も妥当と考えられる。

(b)では仮説を棄却する基準に共通の値 Δ を用いている。しかし、この基準は対象となる2群の関係から設定されるべきものであるため、それぞれ個別に設定する方がよいと思われる。

検出力に関してはさまざまな議論がなされているが、多重比較におけるサンプルサイズを導く理論的な手法は現在確立されていないと思われる。(b)の観点に基づいたサンプルサイズを導くための現実的な手立てとしては、「シミュレーションによる設計」が薦められている[8]。今後の課題としては、(b)の観点に基づいた理論的なサンプルサイズの算出法の導出が挙げられる。

また、(b)において「臨床的に意味のある差」とあるが、薬剤の効果を判定する際に統計手法を用いている以上、「統計学的に意味のある差」も考える必要があると思われる。その2つの「差」に違いがあったとき、サンプルサイズの設計の際にどの「差」を用いるかということも非常に重要な問題であると考えられる。

3.1.2. 複数の評価項目に対する検出力

先述したように、検証的臨床試験の主要評価変数について、ICH-E9では「主要評価変数は通常ただ1つにすべき」と述べられている。ところが、現実には主要評価変数を複数設定し、すべての変数での有効性を示すことが試験の目的となることがある。また、さまざまな疾患に対して効果的な薬が多く開発されていることを考えると、今後複数の変数を評価項目として設定する場面は増加すると考えられる。

このような状況でサンプルサイズを算出する際、変数間の相関を考慮する必要があるが、その場合検出力の計算は複雑なものになる。そのため、現在はシミュレーションによる算出法が紹介されている(「主要評価変数が2つある検証的臨床試験における症例数設計」[7])。今後の課題としては、相関を考慮した複数の評価項目を設定した場合の理論的な症例数算出法の導出が挙げられる。

3.2 複数用量を用いた場合の結果の解釈

薬剤の用量依存性の確認や至適用量の決定をする際、複数用量を用いた用量反応試験が行われる。計画段階での(試験実施者にとって)望ましい結果とは、例えば前者においては用量の増加に従って効果も増加すること(図 3.2 (1))であり、後者においてはある用量まで用量依存性を示し、それを越えた用量に対しては効果の増加が見られないこと(図 3.2 (2))である。しかし実際試験を行うと、思惑から外れた結果が出てくること(図 3.2 (3), (4))。その原因として、主に以下の2点が考えられる。

- ・ 薬剤効果に影響を及ぼす背景(例: 被験者の性別、年齢、疾患の重症度)が投与群間で異なる
- ・ 主要項目がきちんと効果を反映していない(例: 鎮痛剤において、鎮痛作用については用量依存性が見られるが用量の増加に応じて重篤な副作用も発現する場合、主要項目を患者の自己判断から得られる「全般改善度」とすると、原疾患は回復するが副作用が発現したため「全般改善度」の評価が投与前より落ちてしまう)

これらの対処法としては、まず前者については薬剤効果に影響があると考えられる要因を前もって特定した上で、その要因に対して最小化法等を用いることにより各投与群間のバランスをとることが挙げられる。また、後者については試験の目的や薬剤の効果をきちんと反映する指標を主要変数に設定することが考えられる。前述の鎮痛剤を例にすると、まず記入者の主観が入らないように測定値など客観的な指標を用いたり、主観が入らざるを得ない場合には、治療効果として痛みの軽減の度合いのみを抽出するような項目を設定したりすることである。

そのような対処を行ってもまだ解釈しにくい結果が出た場合には、その他の項目(副変数や安全性項目)から得られた結果と合わせて総合的に評価することが必要である。ただし、このような結果からは試験に対する前向きな結論は出せないと思われる。

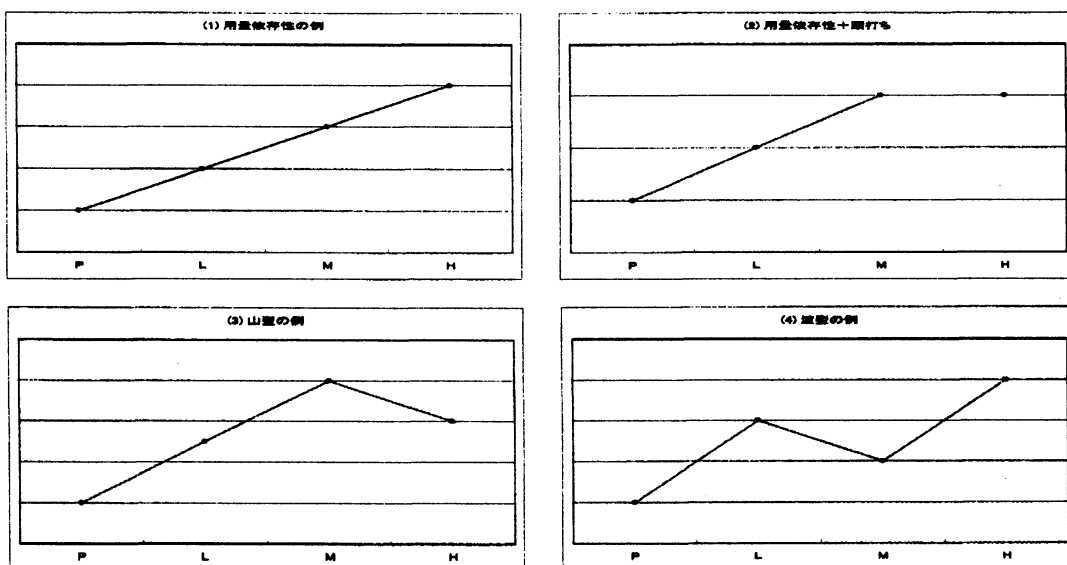


図 3.2 用量反応試験における投与群と効果の関係

3.3 多重性を考慮に入れた信頼区間の構成

現在、臨床試験の解析結果については、P値の他に信頼区間も併記するようになっているが、多重性の調整が必要な場合でも、信頼区間には予め設定した調整していない水準をそのまま用いているのが現状である。しかし、信頼区間もまた「検定」の一種であると考えれば、信頼区間の算出においても多重性の調整を行った水準を用いた方がよいと思われる。

多重性の調整には Bonferroni 法など投与群間の比較を並列的に行う方法もあるが、閉手順など試験の結果によって順序構造を持つような方法も存在する。順序構造を持つ場合、そのことを考慮に入れた信頼区間を算出することが必要になるが、それに関する話題はあまり出ていないようである。順序構造をもつ多重比較における信頼区間の理論的な算出もこれからの課題であると思われる。

<参考文献>

- [1] European Agency for the Evaluation of Medical Products, Committee for Proprietary Medicinal Products (CPMP). Points to Consider on Multiplicity Issues in Clinical Trials. EMEA 2002, <http://www.emea.eu.int/pdfs/human/ewp/090899en.pdf>
- [2] 厚生省医薬安全局審査管理課長. 「臨床試験のための統計的原則」について(医薬審第 1047 号, 平成 10 年 11 月 30 日). <http://www.nihs.go.jp/dig/ich/efficacy/e9/e9j.pdf> 1998
- [3] 厚生省薬務局審査課長. 「新医薬品の承認に必要な用量-反応関係の検討のための指針」について(薬審第 494 号, 平成 6 年 7 月 25 日). <http://www.nihs.go.jp/dig/ich/efficacy/e4/e4.html> 1994
- [4] 藤越慎治. Point to Consider on Multiplicity Issues in Clinical Trials(話題提供者). 2005 年計量生物セミナー, 41-49
- [5] 渡邊裕之, 高橋健一. Point to Consider on Multiplicity Issues in Clinical Trials(討論者). 2005 年計量生物セミナー, 51-54
- [6] 浜田知久馬. 学会・論文発表のための統計学. 1999, 真興交易医書出版部
- [7] 寒水孝司, 叶健. 主要評価変数が 2 つある検証的臨床試験における症例数設計. SAS Forum ユーザー会 学術総会 2004 論文集, 131-140
- [8] 永田靖, 吉田道弘. 統計的多重比較法の基礎. 1997, サイエントリスト社