

2020年6月

多重性の問題が生じる要因【第14回生物統計学】

1. 複数の評価項目を設定する場合

複数の評価項目を設定することは、検定の仮説が増えることから多重性の問題が生じる。

- 複数の主要評価項目を設定する場合
主要アウトカムを複数設定する場合、次の2つのパターンが考えられます。

- ①全てを検証する必要がある場合 (co-primary endpoint)
- ②いずれかの項目が検証できれば良い場合 (multi-primary endpoint)

①と②では多重性の調整方法が異なるので、以下に紹介します。

<co-primary endpoint>・・・X個の主要評価項目のうち、いずれも検証する必要がある場合は、X個の検定で同時に有意差を検出できた場合のみ試験の目的が達成される。この場合、多重性の問題は生じない。しかし、検出力の低下に注意する。

<multi-primary endpoint>・・・X個の主要評価項目のうち、いずれかを検証すれば良い場合は、第1種の過誤確率が上昇するため、多重性の問題が生じるため、有意水準または有意確率の調整が必要になる。

- 主要副次評価項目（重要な副次的評価項目）を設定する場合
主要評価項目に加えて、更に機能性表示食品の付加価値の付与などを目的として、副次評価項目の中でも検定仮説を設定することがあると思います。この場合、上述の multi-primary endpoint に類似した状況であるため多重性の問題が生じますが、複数の評価項目の重要度が同等では無く、明らかな順序性があるといった違いがあります。原則として、主要評価項目が検証された場合に主要副次評価項目が解釈可能である事になるので注意が必要です。
- 複合評価項目を設定する場合
機能性表示食品の臨床試験では、現時点では採用されておりませんが、複合評価項目というものがあります。第12回生物統計学でも紹介した ICH-E9 に定義が記載されているので参照してください（合成変数として説明されています）。主要な目的に関する複数の評価指標の中から、主要評価項

目として1つを選ぶことができない場合などに事前に複数の評価指標を合成して単一の評価項目を定義することも多重性の問題に対応する選択肢の1つとされています。複合評価項目自体を1つの主要評価項目として解析する場合は、多重性の問題は生じませんが、それを構成する複数の評価指標について主要副次評価項目の様に見なす場合は多重性を調整する必要があります。

2. 三つ以上の群を設定する場合

三つ以上の群が設定され、対比較を複数回行う場合にも多重性の問題が生じます。

群の構成と仮説における想定と検定手順（例えば比較群間で有効性に順序があることを想定できる場合や1つの対照群とその他の群の対比較を想定する場合など）によって多重性調整法が異なります。また、群間比較の回数が多くなればなるほど第1種の過誤確率が上昇するので多重性を考慮すると有意差を検出しにくくなります。そこで、相対的に重要度が低いと考えられる対比較（例えば、用量設定試験における低用量群 vs 高用量群のような比較）は計画しないという対策を考えられます。

3. 部分集団の解析を実施する場合

健常者と軽症域者が混在している際に、健常者と軽症域者で集団を形成して、解析することがありかと思えます。そのような解析を部分集団解析と呼びます。これは、主目的達成後の補強または仮説生成を目的とした探索的な目的、検証的な目的で実施されます。例えば、全体集団における有効性を検証し、次に部分集団における有効性の検証を行う場合、検定を複数回実施する必要が生じるため多重性の問題が生じます。

4. 評価時点が複数ある場合

主要な解析（主解析）に用いられる評価時点として例えば12週時など特定の1時点が用いられることがあります。12週時だけの群間比較のみを行う場合は多重性の問題は生じません。一方で、12週時の測定を最終評価時点とした場合、4週、8週と測定していることが多く、評価時点が複数あり、経時的に介入効果を検討する場合があります。この検討において時点毎に群間比較の検定を繰り返した場合（いわゆる輪切りの検定）は多重性の問題を生じます。

5. 中間解析を実施する場合

中間解析は、臨床試験が正式に完了する前に行われる有効性または安全性に関する群間比較を意図したすべての解析を指します。これは、医薬品の開発において上市までの効率を高めるためのアダプティブデザインなどに用いられるもので、機能性表示食品の臨床試験で中間解析を適用することは現時点ではありません。中間解析を行うたびに有意水準が小さくなる特徴があります。

6. 同じ評価項目へ複数の解析手法を適用する場合

1つの評価項目に対して複数の検定方法を適用する場合も多重性の問題が生じます。例えば、12週時のBMIに対して student の t 検定と Wilcoxon 順位和検定の両方を適用する場合などです。事前情報がない場合、2種類の解析結果を見てから判断したいと考えたいと思いますが、このような場合は多重性の問題が発生しているので注意して下さい。この事例のように1つの結果を導くために複数の検討を基にする場合（いいとこどりの状況）は、多重性の問題が生じます。

7. 信頼区間を複数報告し解析手法の結果のように解釈する場合

信頼区間は、介入効果の大きさの信頼度を定量的に示す有益な指標であり、常に介入効果の推定値と共に報告されることが望ましいとされています。信頼区間は、多くの検定方法において $P < 0.05$ の場合に介入効果の95%信頼区間が0（差の場合）もしくは1（比の場合）を含みません。「差」の指標では「信頼区間が0を含むか否か」、「比」の指標では「信頼区間が1を含むか否か」とそれぞれに対応する「検定が有意であるか否か」の間には1:1の対応関係があります。よって、複数の検定を実施しなくても、信頼区間を複数算出し、それぞれについて多重性への考慮なく帰無仮説を含んでいるか否かを判定すると検定を一度も実施せずとも多重性の問題が生じます。

8. まとめ

今回は多重性の問題が生じる状況をまとめました。多重性の問題は、機能性表示食品のガイドラインでも重要視されています。そのため、プロトコルを作成する段階で、仮説設定や評価項目の設定を念密に行う必要があるため、身近な生物統計家に相談してください。

臨床試験は是非、オルトメディコまで!