臨床試験における多重性の諸問題の現状と今後の課題について

―非統計家向け解説書―

日本製薬工業協会 医薬品評価委員会

データサイエンス部会 2019年度
タスクフォース 4 多重性調整チーム

Ver 1.0

2020 年　6月

目次

＜図の一覧＞

＜表の一覧＞

＜付録の一覧＞

＜用語＞

用語	内容
labeling claim	添付文書（米国における labeling および欧州（中央審査）における Summary of Product Characteristics; SmPC）に記載する効能・効果の範囲

＜略語＞

略語	英名	和名
ACR	American College of Rheumatology	米国リウマチ学会
AUC	Area under the curve	局面下面積
CHMP	Committee for Medicinal Products for Human Use	欧州医薬品委員会
CI	Confidence interval	信頼区間
COPD	Chronic Obstructive Pulmonary Disease	慢性肺閉塞性疾患
CPMP	Committee for Proprietary Medicinal Products	欧州医薬品委員会
DAS	Disease Activity Score	疾患活動性スコア
EMA	European Medicines Agency	欧州医薬品庁
FDA	Food and Drug Administration	アメリカ食品医薬品局
FEV_1	Forced Expiratory Volume in the first second	努力性肺活量のうちの最初の 1 秒間に吐き出した空気の量
FWER	Family-wise Error Rate	仮説族における第 1 種の過誤確率
HAM-D	Hamilton Rating Scale for Depression	ハミルトンうつ病評価尺度
HAQ-DI	Health Assessment Questionnaire-Disability Index	日常生活機能
hsCRP	High sensitivity C-reactive protein	高感度 C-反応性タンパク質
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use	医薬品規制調和国際会議
ITT	Intent-to-treat	治療に用いる治療方針により得られる効果は，実際に受けた試験治療ではなく，被験者を治療しようとする意図（予定した試験治療規定）に基づくことにより最もよく評価できる，ということを主張する原則
mTSS	Modified total Sharp score	修正総合シャープスコア
OS	Overall Survival	全生存期間
PANSS	Positive and Negative Syndrome Scale	陽性・陰性症状評価尺度
PFS	Progression-Free Survival	無憎悪生存期間
PMDA	Pharmaceuticals and Medical Devices Agency	独立行政法人医薬品医療機器総合機構

QOL	Quality of Life	クオリティ・オブ・ライフ
SDAI	Simplified disease activity index	疾患活動性評価
SGRQ	St. George's Respiratory Questionnaire	セント・ジョージ式呼吸器質問票
TDI	Mahler Transitional Dyspnea Index	息切れの評価指標

1. 序文

1.1. エグゼクティブ・サマリー

医薬品開発における統計的仮説検定では，本来は効果がない被験薬を「効果あり」と判断する誤りと，本来は効果がある被験薬を「効果なし」と判断する 2 つの誤りが存在する。臨床試験において統計学的仮説検定で判断を下す場合には，被験者数を大きくすることなくこれら 2 つの誤りを同時に小さくすることが出来ない。しかし，効果がない薬剤を世に出す確率，すなわち消費者危険を大きくすることはできないので，効果がない被験薬を「効果あり」と判断する確率を許容できるレベル以下に抑えることが優先されるため，医薬品の承認申請を目的とした検証試験では規制上，消費者危険の厳密な制御が必要となる。この制御の基準がいわゆる統計的仮説検定における有意水準であり，慣例的に 5%とされる（ICH ,1998)。

検定を複数回実施し，かつ意思決定の選択肢が複数存在する [1]と，本来は「効果がない」被験薬であっても，誤って「効果がある」と判断してしまう確率が有意水準を超過することが知られている。これを検定に伴う多重性の問題という。多重性の問題を引き起こす要因は様々知られており，複数の評価項目の比較，3 群以上の投与群の比較，複数の評価時点毎の比較，中間解析のような複数の解析時点における比較などがあり，臨床試験の様々な場面で検定に伴う多重性の問題を引き起こす要因が存在する。なお，検定を複数回実施したとしても，全ての検定が有意であったときにのみ試験成功とみなす場合，すなわち試験成功の定義が 1 通りである場合においては，多重性の問題が生じない点については注意する。

多くの場合，検定結果と信頼区間が帰無仮説の値を含むか否かの結果には対応関係があることが知られている。例えば「推定されたハザード比が有意か否か」と，「ハザード比の信頼区間が帰無仮説（すなわちハザード比＝1）を含んでいるか否か」には 1 対 1 の対応関係がある。このため，検定を実施せずとも信頼区間を複数構成し，帰無仮説を含んでいるか否かをあたかも検定を実施しているかのように用いると，検定は実施していなくとも多重性の問題が生じうる点には注意が必要である。

医薬品に対する差別化および付加価値の付与の目的や，規制要件など踏まえ，複数の主要評価項目が設定され，更に主要副次評価項目が設定されることがある。この場合には，複数の検定仮説が存在することから，検定の順序や，意思決定の枠組み（例えば，1 つでも有意な結果があれば試験成功とするのか [2]，複数の項目で有意な結果が得られないと試験成功と見なさないかなど）を予め考慮しておく必要がある。検定と意思決定の妥当な枠組みは，薬剤の位置づけ，治療ガイドラインや各規制当局の方針などによって影響を受ける。承認申請後に規制当局から本枠組みの変更を指示された事例も存在することから，試験計画段階でその妥当性を規制当局と合意しておくことが望ましい。

検定に伴う多重性の問題についての統計学的な対処法については，これまでに多くの研究がなされ，様々な多重性の調整法が既に利用されている。従来はボンフェローニ法など，比較的簡便な多重性調整方法が使用されていたものの，仮説構造の複雑化やより検出力の高い方法への需要の増加に伴い，ゲートキーピング法に代表されるような，より複雑な仮説構造に対応した多重性調整方法が提案されてきた。さらに，グラフィカル・アプローチのような多重性調整の方針を関係者で分かり易く共有する方法も提案され，実際の医薬品開発

[1] 例えば，多重性の調整をせず複数回の検定を実施した時，いずれか 1 つでも有意となった場合に試験成功と見なす事例等が該当する。

[2] 上述のように，この場合は多重性の調整が必要になる。

においても適用事例が報告されている。これら多重性調整方法の高度化に伴い，試験目的および仮説構造に合致した適切な方法を用いるためには方法の深い理解が必要となり，高度な専門性が要求される。

目的に合致した妥当な仮説構造とそれに伴う意思決定の枠組みの議論は，統計解析担当者のみならず統計解析担当者以外の多くの関係機能に関わる極めて重要な課題である。意思決定の透明性の観点から，統計解析担当者は医薬品開発に関わる様々な関係者に対し，仮説構造および意思決定の方法や結果とその解釈をわかりやすく説明するコミュニケーション能力が要求される。

検証試験ではない探索的位置づけの試験においては，検定に伴う多重性の厳密な制御は規制上要求されない。しかし，十分に検出力を確保した試験計画であっても，明確な意思決定の枠組みを持たず，多重性の調整をおざなりにした場合には，偶然認められた結果（例えばいわゆるチャンピオンデータ）に基づきその後の開発を計画する可能性があることから，探索的試験であっても多重性調整の要否には慎重な議論が必要である。

今後は本邦においても複雑な仮説構造を伴う試験が増加する可能性がある。これら医薬品開発を取り巻く環境の変化から，仮説構造とそれに伴う意思決定の方法を含めた試験計画の策定は，統計解析担当など一部の臨床開発担当者のみの問題に留まらず，マーケティングや薬事を初めとした多くの関係機能に亘る課題となるため，開発計画の策定並びに試験実施計画書の作成を開始する早期の段階から関係機能と議論を開始する必要がある事に留意すべきである。

1.2. 本稿の構成

第 2 章から第 6 章までの構成は以下の通りである。第 2 章では，検定に伴う多重性を理解するにあたり，検定の枠組みについて解説し，多重性を生じさせる要因について触れる。第 3 章では，ICH および各規制当局の多重性に関する立場について，共通点・相違点をまとめる。第 4 章では，多重性の調整方法が審査において議論となった事例や，複雑な仮説構造など興味深い多重性調整の事例を紹介する。第 5 章では，多重性の問題を考えるにあたっての今後の課題について触れる。

本稿は，統計を専門としない医薬品開発関係者への検定に伴う多重性の問題についての解説を意図しているため，多重性調整の方法論についての解説は割愛した。必要に応じて多重性調整方法について解説した総説（Dmitrienko et al., 2013; 寒水, 2015）や成書（Dmitrienko et al., 2009; 坂巻ら, 2019）などを参照されたい。

多重性の調整方法に関しては有意水準を調整する立場と，検定結果の p 値を調整する立場に大別されるが，本稿では特に断りがない限り有意水準を調整する立場をとる。

本稿が医薬品開発関係者の検定に伴う多重性について理解を深め，適切な医薬品開発の一助となれば幸いである。

【参考文献】

[1] ICH. (1998). ICH harmonised tripartite guideline: statistical principles for clinical trials-E9. Available online at: https://database.ich.org/sites/default/files/E9_Guideline.pdf

[2] Dmitrienko, A., & D'Agostino Sr, R. (2013). Traditional multiplicity adjustment methods in clinical trials. Statistics in Medicine, 32(29), 5172-5218.

[3] 寒水孝司. (2015). 臨床試験における多重性の諸問題. 計量生物学, 36(Special_Issue_2), S87-S98.

[4] 坂巻顕太郎, 寒水孝司, & 濱崎俊光. (2019). 多重比較法. 朝倉書店.

[5] Dmitrienko, A., Tamhane, A. C., & Bretz, F. (Eds.). (2009). Multiple testing problems in pharmaceutical statistics. CRC press.

2. 多重性とは

2.1. 検定における 2 種類の誤り

医薬品開発において被験薬に効果がある事を証明しようとする場合，まず「被験薬と対照薬（プラセボなど）の治療効果に差がない」という仮説（帰無仮説）を立て，帰無仮説を否定することで「被検薬と対照薬の治療効果に差がある」という仮説（対立仮説）を採択し，「被験薬に治療効果がある」と判断する手順が取られる。この一連の手順を統計的仮説検定（以後，検定）と呼ぶ。回りくどい方法に思えるが，論理的には「治療効果に差がある」ことを直接証明することは不可能であるため，検定では背理法に基づく上記の手順を用いる。

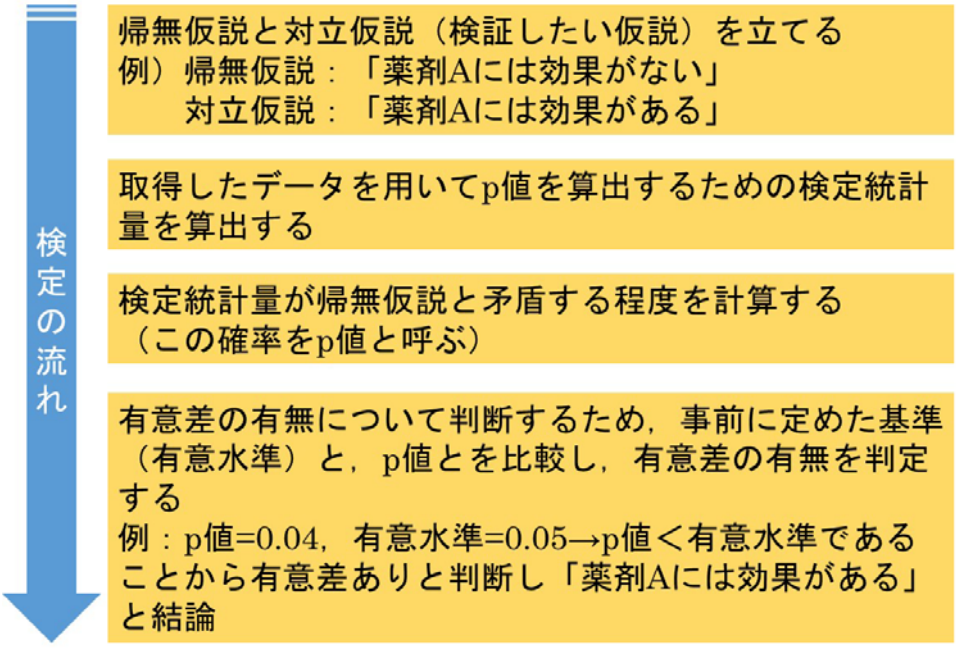


図 2.1-1　統計的仮説検定の流れ

治療効果があるか否かについての真実は分からない下で検定を用いて意思決定を行うため，統計的仮説検定では誤った判断をする可能性が常に存在する。特に医薬品開発の枠組みでは，下記の 2 種類の誤りが存在する。

- ✓ 本来は効果がない被験薬を，効果があると判断する誤り（第 1 種の過誤）
- ✓ 本来は効果がある被験薬を，効果がないと判断する誤り（第 2 種の過誤）

表 2.1-1　検定における 2 種類の過誤

		検定結果	
		有意差あり	有意差なし
真実	薬効なし	α（第 1 種の過誤）	1－α
	薬効あり	1－β（検出力）	β（第 2 種の過誤）

いずれも避けるべき誤りであるが，第 2 種の過誤は治療機会の損失にはなるものの，現状に変化を生じさせない事に対し，医薬品は安全性上のリスクが存在するという立場の下では，第 1 種の過誤は誤った治療選択肢を増やすことで消費者危険を増大させ，社会的損失を招く恐れがある。また，効果がない薬剤が投与されることに起因する本来可能な治療機会の損失にもつながる。図 2.1-2 に示す通り，これら 2 種類の過誤はいずれも同時に小さ

くすることは出来ないため，消費者危険を重視する立場から第 1 種の過誤を許容可能な範囲まで小さくする事が求められる。その考え方に基づき，International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use（医薬品規制調和国際会議）の E9 ガイドラインである“臨床試験のための統計的原則”（1998）（以後，ICH E9 と呼ぶ）では，検証試験においては第 1 種の過誤を 5%以下，第 2 種の過誤を 10～20%に抑えることとされている。

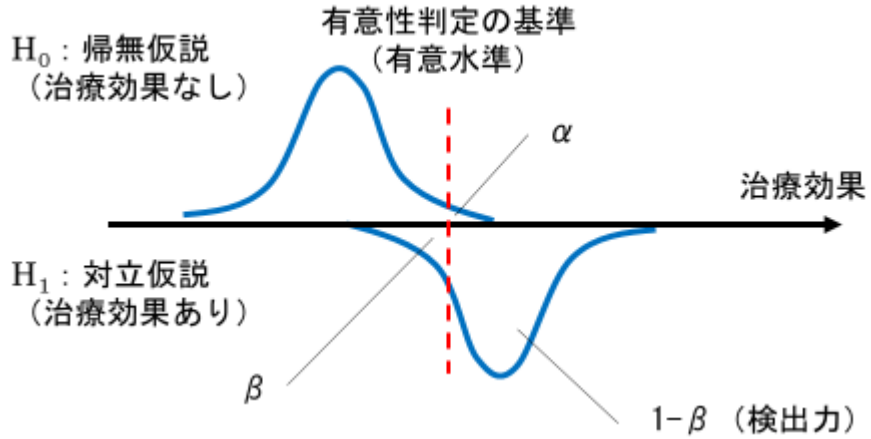


図 2.1-2　帰無仮説および対立仮説の下での検定統計量の分布と 2 種類の過誤

釣鐘型の分布はそれぞれの仮説における想定される治療効果の分布を示す。帰無仮説（治療効果なし）の下での赤の波線（有意性判定の基準）の右側の面積が第 1 種の過誤確率を示す。一方，対立仮説（治療効果あり）の下での赤の波線の左側の面積は第 2 種の過誤確率を示す（右側の面積は検出力に対応する）。各仮説を定めた下では，分布の位置は固定されるため，赤の波線を右にずらせば第 1 種の過誤を小さくできるが，第 2 種の過誤は増大し（検出力が小さくなる），一方，赤の波線を左にずらすと第 2 種の過誤は小さくなるが，第 1 種の過誤は増大する。これら 2 種類の過誤は同時に小さくすることが出来ないことが分かる。

2.1.1. 検定に伴う多重性の問題

検定では，帰無仮説が正しい下で誤って対立仮説を採択する誤り，すなわち「本来は差がない」にも係わらず差があると判断してしまう誤りの確率（正しい帰無仮説が誤って棄却されてしまう確率）を第 1 種の過誤確率と呼ぶ。ICH E9（1998）では，慣例的に検証的試験における第 1 種の過誤確率を 5%以下に設定することとされている。第 1 種の過誤確率として許容できる閾値を，有意水準と呼び，検定の際に算出される p 値の閾値として用いられ，p 値が 0.05 未満であれば，「治療効果に有意差あり」と判断される。

複数回検定を実施することにより，検定に伴う多重性が発生するメカニズムを説明するにあたり，2 つの評価項目に対する検定のうち，少なくともいずれか一方で有意差が認められた時に試験成功とする，つまり，被験薬に効果があると判断する場合を考える。本来は薬効がない下で検定を実施した場合に取りうる検定結果は下記の 4 通りとなる（図 2.1.1-1）。

- ✓ 1 つ目の検定のみが誤って薬効ありとする場合：①
- ✓ 2 つ目の検定のみが誤って薬効ありとする場合：②
- ✓ いずれの検定も誤って薬効ありとする場合：③
- ✓ いずれの検定も正しく薬効なしとする場合：④

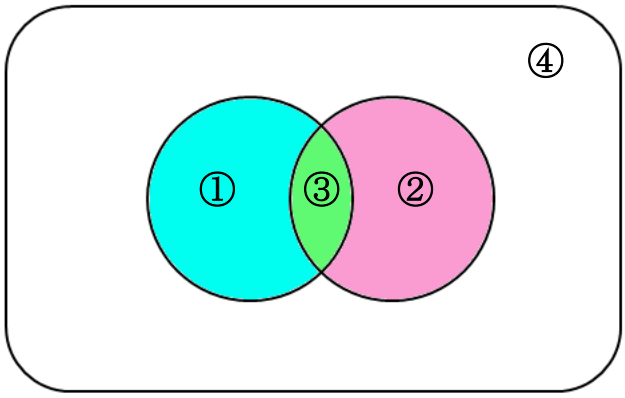


図 2.1.1-1　帰無仮説が正しい（被験薬と対照薬の治療効果に差がない）下で検定を 2 回実施した際の第 1 種の過誤確率の概念図

この時,「薬効がない」と正しく判断できているのは④であり,「薬効がある」と誤って判断しているのは①，②，③の場合となる。先に述べたとおり,「薬効がない」被験薬を「薬効がある」と判断してしまう確率を 5%以下に抑える必要があるが，2 つの円のそれぞれが 5%となるように正しい検定方法を用いたとしても，①＋②＋③の領域の合計は 1 つの円の大きさを超過しているため，5%を越えていることになる。これにより，個々の検定で有意水準が 5%となるように検定を実施したとしても，検定を複数回実施することによって，全体の第 1 種の過誤確率が上昇することがわかる。本稿ではこの現象を「検定に伴う多重性の問題」と呼ぶ。

2 つの検定に関連性がない独立の場合，④の確率は 0.95×0.95＝0.9025（90.25%），①＋②＋③の確率は 1－0.9025＝0.0975（9.75%）となり，本来は「薬効がない」薬を「薬効あり」と判断してしまう誤りとして許容した 5%に対し，9.75%は約 2 倍となっており，検定を 2 回実施したことにより多重性の問題が生じていることがわかる。

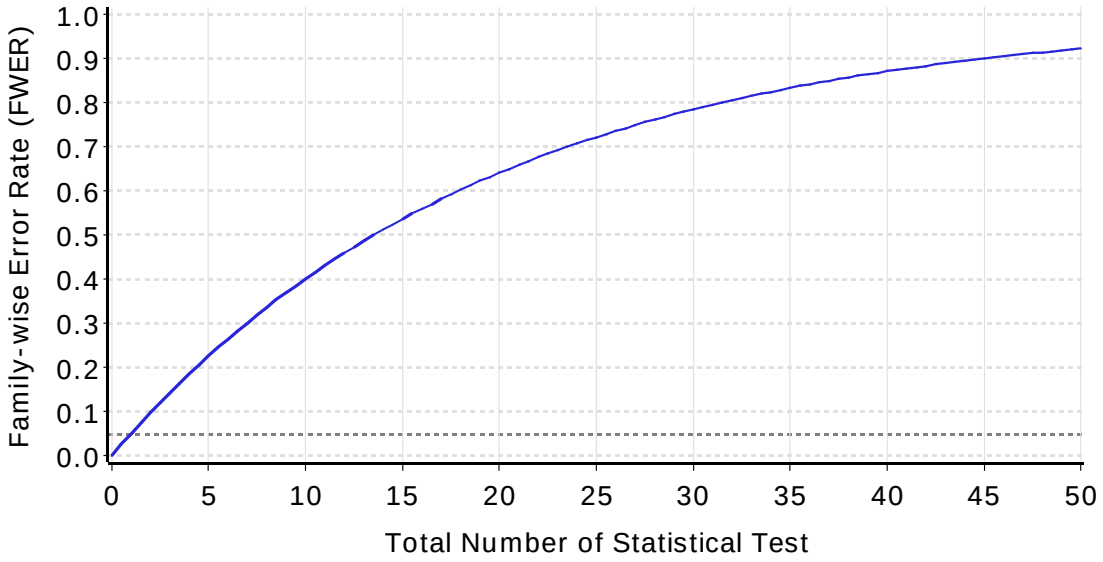


図 2.1.1-2　検定回数と全体の第 1 種の過誤確率（FWER）の関係

個々の検定は有意水準 0.05（5%）で実施し，互いに独立であると仮定。検定を多数実施することで，個々の検定は所定の有意水準であっても全体の第 1 種の過誤確率（FWER）が増大することが分かる。検定を

50 回実施すると，いずれの仮説も本来は差がないにもかかわらず，有意差ありと判断する確率が約 90%にもなる。

検証すべき仮説の集まりは「仮説族」や「ファミリー」と呼ばれ，それらの仮説の中で差が無いにも係わらず，少なくとも 1 つの検定で差があると判断してしまう確率を Family-wise error rate（FWER）と呼ぶ。上記の例では，FWER が 9.75%となる。先に述べたように検証試験では FWER を厳密に 5%以内に制御することが求められる。図 2.1.1-2 に検定回数と FWER の関係を示した。検定の回数が増えるにつれ FWER が増加し，検定回数が 10 回では FWER は約 40%，50 回ともなると FWER は 90%を超過する。

FWER を 5%に制御するための多重性調整法として，様々な方法が提案されている。これら方法については，多重性調整方法についての総説（Dmitrienko et al., 2013; 寒水, 2015）や成書（Dmitrienko et al., 2009; 坂巻ら, 2019）を参照されたい。

2.1.2. 第 2 種の過誤確率の制御と検出力の確保

検定を行う際に，真実は「差がある」にも係わらず，差がないと判断してしまう確率を第 2 種の過誤確率と呼ぶ。一方，真実は「差がある」場合に，正しく「差がある」と判断する確率を検出力と呼ぶ。定義上，第 2 種の過誤確率と検出力は合計すると 1（100%）になる。ICH E9（1998）では，慣例として検証的試験の第 2 種の過誤確率を 10～20%に設定する旨が記載されている。第 1 種の過誤確率と第 2 種の過誤確率は，いずれか一方が減少すると他方が増加するという関係がある（図 2.1-2）。ただし，被験者数（生存時間解析においてはイベント数）を増やせば，図 2.1-2 における分布のばらつきがより小さくなることから，いずれの過誤確率も同時に減少させることが出来る。しかし臨床試験の倫理的側面から，不必要に多くの被験者を試験に組み入れることは出来ないため，許容可能な第 1 種の過誤確率および第 2 種の過誤確率（検証的試験では，第 1 種の過誤確率が 5%，第 2 種の過誤確率が 10～20%）に基づき必要被験者数を算出することになる。

ここで，図 2.1.1-1 と同様の状況を考える。ただし，図 2.1.2-1 は対立仮説が成立する条件下，つまり本来は薬効がある被験薬（いずれの帰無仮説も正しくない）に対しての概念図になっている。図 2.1.1-1 と同様に，下記の 4 通りの結果が想定される。

- ✓ 1 つ目の検定のみが正しく薬効ありとする場合：①
- ✓ 2 つ目の検定のみが正しく薬効ありとする場合：②
- ✓ いずれの検定も正しく薬効ありとする場合：③
- ✓ いずれの検定も誤って薬効なしとする場合：④

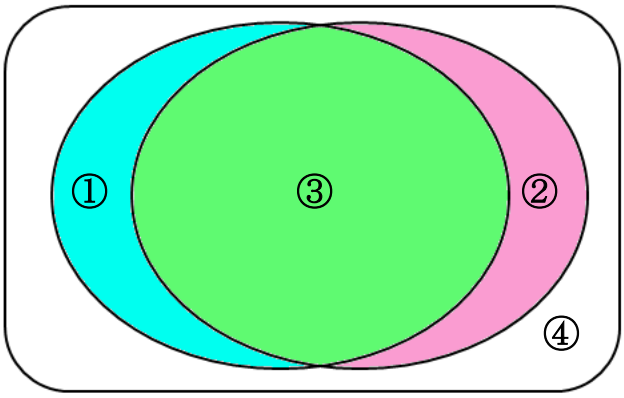


図 2.1.2-1　対立仮説が正しい下で検定を 2 回実施した際の検出力の概念図

①と③で構成される円は 1 つ目の検定の検出力を示しており，②と③で構成される円は 2 つ目の検定の検出力を示す。円が重複する③の部分は双方で薬効を正しく判断できる状態であり，円の外側である④の部分はいずれの検定でも薬効の判断を誤ってしまう状態を示す。この時，いずれかの検定で「薬効がある」と正しく判断できるのは①，②，③であり，いずれも「薬効がない」と誤って判断してしまう場合は④となる。ただし，①および②では一方の仮説を正しく判断できていないことから，両方の仮説を同時に検証したい場合は，③の部分が一定の検出力（例えば 80～90%）を保てるように症例数を設定する必要がある。

なお，各検定における評価項目間に関連がなく独立であり，それぞれの検定の検出力を 80%とする場合，③の確率は 80%×80%＝64%となり，2 つの検定に対する検出力は個々の検定の検出力より小さくなることに注意が必要である。2 つの検定における検出力を 80%程度確保したい場合は，個々の検定の検出力を 90%とすることにより，90%×90%＝81%となることから，考慮することができる。この場合は，個々の検定の検出力を 90%に上げる必要があるため，必要被験者数（あるいはイベント数）が多くなる。

2.2. 多重性を生じさせる要因

2.2.1. 評価項目が複数ある場合

ICH E9（1998）では，「主要評価項目を通常ただ 1 つにすべきである」と述べられており，その場合は評価項目が複数存在することによる多重性の問題は生じない。しかしながら，特定の疾患領域では複数の主要評価項目を検証する必要があり，また，labeling claim などの添付文書の効能・効果および臨床試験成績の記載へ反映するなどの目的で主要副次評価項目を設定する場合は，検定を実施する評価項目が複数となるため多重性の調整が必要となる。

✓ 複数の主要評価項目

複数の主要評価項目を設定する場合において，「①その全てを検証する必要がある場合（co-primary endpoint）」と，「②いずれかの項目が検証できれば良い場合（multi-primary endpoint）」では，多重性の調整の要否が異なる。以下では簡単のために主要評価項目が 2 個の場合で記載しているが，3 個以上でも考え方は変わらない。

- co-primary endpoint

  2 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，2 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 [3]。この場合，FWER は図 2.1.1-1 の③に該当し，各検定を 5%で行う場合，明らかに 5%を下回っているため，多重性の問題は生じない。しかし，2.1.2 で触れたように，検出力が低下する点には注意する。

- multi-primary endpoint

  2 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，図 2.1.1-1 における①～③の和が FWER に該当する。この場合，各検定を 5%で行う場合，FWER は明らかに 5%を上回り多重性の問題が生じることから，多重性の調整が必要になる。

✓ 主要副次評価項目

試験の成功可否を判断するための主要評価項目に加え，更に薬剤の付加価値の付与など

[3] 注）慣例的に主要評価項目を 2 つ設定した場合（例：PFS と OS）に co-primary endpoint と呼ぶ事例も存在する。

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。上記の multi-primary endpoint に類似した状況であり多重性の問題が生じるが，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る。

原則として，主要評価項目が検証された場合にのみ，主要副次評価項目が解釈可能である事には注意が必要である。

✓ 複合評価項目

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 つである。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，多重性の問題は生じないが，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には，多重性を調整する必要が生じる。

複合評価項目を用いる場合には，多重性の問題以外にも注意すべき点がある。例えば，複数のイベント（入院，死亡など）の転帰によって複合評価項目が構成される場合[4]は，定義されたいずれかのイベントが生じた場合に複合評価項目としてイベントが生じたと見なされることになるが，それぞれのイベントの重要性が異なる場合がある。重要性が高い死亡のようなイベントと重要性が低い他のイベントが構成要素に含まれている場合，死亡では対照薬に劣っていても，重要性が低い他のイベントで勝る場合に，複合評価項目としては対照薬に勝ると判断する可能性がある。

状況によっては結果全体の解釈に影響を及ぼす可能性があるため，複合評価項目の妥当性を補足する目的において構成要素の解析は重要である。

一方で，PANSS や HAM-D といった精神症状評価尺度や QOL 評価尺度などでよく見られるように，複数の質問項目の合計スコアを評価指標とする場合[5]，合計スコアは臨床的に解釈可能と見なされるが，下位スコアあるいは下位ドメインの中にはそれのみでは臨床的重要性が不明で，解釈できないものが存在する尺度がある。そのため，臨床的重要性が不明で解釈困難と考えられる下位スコアの解析では，全体的な結果にいずれの下位スコアが最も寄与しているかを理解する目的で有益な場合があるが，あくまで補足的に使用すべきであり，結論を誇張しない方法で要約統計量などを用いて記述的な分析結果を提示することが重要である。

2.2.2. 投与群が複数ある場合

3 群以上の投与群が設定され，検定による対比較（いずれかの 2 群間での比較）を複数回行う場合，多重性の問題が生ずる。投与群には，被検薬の複数用量やプラセボ群，実薬対照群のような対照群が想定され，それらの群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る。

2 群間の検定による比較の回数が多くなればなるほど，多重性の処理が困難になることがあるため，相対的に重要度が低いと考えられる実薬低用量群 vs 高用量群のような比較は

[4] FDA (2017)では”composite endpoint”と呼ばれている。
[5] FDA (2017)では”multi-component endpoint”と呼ばれている。

行われないことがある。最近では，多重性の調整を行った上で用量反応関係（用量反応曲線の形状など）の検討を行うことができる MCP-Mod（Bretz et al., 2005）を用いた報告もある。

2.2.3. 部分集団解析の実施

部分集団解析は主目的が達成された後にそれを補強する，もしくは，仮説生成を目的とした探索的な文脈で実施される。一方で，部分集団解析が検証的な目的で実施されることもある。例えば，事前に特定の部分集団で効果が異なることが示唆されており，labeling claim への記載追加を意図している時，まず全体集団における有効性を検証し，次に当該部分集団における有効性の検証を行うことがある。その場合，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る。

バイオマーカーを用いた医薬品開発においても，部分集団解析における多重性が問題となる場合がある。バイオマーカー陽性の部分集団で薬効が望まれる，あるいはより強く効果を示すことが期待されるなど，バイオマーカーが患者選択に用いられる場合，バイオマーカー陽性の部分集団のみで試験を実施し，承認取得を計画したとしても，陰性集団も試験に組み入れることが推奨されることがある（FDA, 2019）。これに従い，バイオマーカー陰性の部分集団も試験に組み入れ，陽性の部分集団の有効性の検証を行った後に，陰性も含めた全体集団で効果の検証を行う場合は多重性の問題が生じる。なお，全体集団で統計的な有意差が認められたとしても，陰性集団における探索的な検討を踏まえて，陰性集団が claim から除外されることもあり得る（Johnston et al., 2009）。

他方で，多数の部分集団解析を探索的に行うことによる多重性の問題がある。仮説生成を目的とした探索的な部分集団解析であれば，必ずしも多重性を調整する必要はないが，明確な意思決定の枠組みを持たず漫然と実施した場合，誤った意思決定を行う危険性がある事に留意されたい。

2.2.4. 評価時点が複数ある場合

主要解析に用いられる評価時点として，例えば投与後 24 週時など，特定の 1 時点が用いられることがある。ただし，臨床試験では投与開始から投与終了までの間に複数の Visit が設定されることが多く，経時的な薬効の検討を行うことがある。この検討に際して，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る。

2.2.5. 中間解析を実施する場合

中間解析とは，試験が正式に完了する前に行われる有効性又は安全性に関する試験治療群間の比較を意図したすべての解析を指す（ICH E9, 1998）。特に，有効中止を目的とした中間解析を実施する場合，試験完了時の最終解析も含めて複数回の検定を実施することになるため，多重性の問題が生じる。本稿の範囲を超えるため詳細については触れないが，有効中止を目的とした中間解析における多重性の調整にはα消費関数を用いた方法（Lan-DeMets 法）が用いられることがある。α消費関数は，中間解析を実施する度に文字通りαを消費し，最終解析時に使用できる（残された）有意水準を求めることができる方法である。よって，有効中止を目的とした中間解析を実施する際には，最終解析時の有意水準は事前に規定した有意水準（例えば 5%）よりも小さな値となるため，中間解析では試験が中止されず最終解析に到達した場合は，有意な結果を得るためにはより小さい p 値となる必要がある。α消費関数には，幾つかの種類があるため，試験の目的に応じて適切な方法を選ぶ必要

がある。

2.2.6. 同一評価項目へ複数の解析方法を適用する場合

連続量の評価項目に対して，パラメトリックな方法である「対応のない t 検定」と，ノンパラメトリックな方法である「Wilcoxon 順位和検定」の双方で解析したとする。当該評価項目に対する解析方法としていずれが妥当であるかは，母集団の分布に依存する。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ることがあるかもしれないが，このような場合は多重性の問題が発生している。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，すなわち，「いいとこどり」の状況では，多重性の問題が発生する。

2.2.7. 信頼区間を複数報告し検定結果のように解釈する場合

治療効果の推定値の信頼性を示す尺度として信頼区間がある。信頼区間そのものは治療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，常に治療効果の推定値と共に報告されることが望ましい。

信頼区間は検定と表裏一体の性質があり，多くの検定方法においては，P<0.05 の場合に治療効果の 95%信頼区間が 0（差の場合）もしくは 1（比の場合）を含まないことが知られている [6]。「治療効果の検定結果が有意か否か」と，「その信頼区間が帰無仮説を含むか否か」には表裏一体の関係がある。つまり，リスク差など「差」の指標について考えている場合には「信頼区間が 0 を含むか否か」，オッズ比やハザード比など「比」の指標について考えている場合には「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 の対応関係がある。

このため，複数の検定を実施しなくとも，その代わりに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る点には注意が必要である [7]。

2.3. 探索的試験における多重性の問題

探索的試験では，様々な側面から統計解析が行われ，事後的な追加解析が行われることも少なくはないが，規制上は多重性を調整する必要は無いとされている。しかしながら，多重性を調整する必要が無いことと多重性の問題が無いことは同義では無い。探索的な試験においても，明確な意思決定の枠組みを持たず多重性の問題を考慮しないと，多くの探索解析の中から見出された結果に基づきその後の開発の意思決定を行う可能性が生じるため，探索的試験であっても多重性の問題は意思決定に誤りを生じさせる可能性があることに注意

---

[6] 注：PFS や OS などイベント発現までの時間を評価項目とした場合，主要評価項目の検定はログランク検定，治療効果の推定値（ハザード比）には Cox の比例ハザードモデルが用いられる。医薬品開発における生存時間解析では検定と推定において，慣例的に異なる方法を用いるため，検定結果と信頼区間は 1 対 1 に対応しない。

[7] 補足：信頼区間にも多重性を調整する方法は提案されているが，多重性の調整方法によっては検定に対応する多重性を調整した信頼区間を構成することが困難な場合がある。その場合に，検定結果と合わせて信頼区間を提示するためには，いずれも検定結果と対応関係は持たせられないが，多重性を調整しない信頼区間もしくはより保守的な多重性調整方法（例：ボンフェローニ法）を用いることが考えられる。

が必要である。

【参考文献】

[1] Dmitrienko, A., & D'Agostino Sr, R. (2013). Traditional multiplicity adjustment methods in clinical trials. Statistics in Medicine, 32(29), 5172-5218.
[2] 寒水孝司. (2015). 臨床試験における多重性の諸問題. 計量生物学, 36(Special_Issue_2), S87-S98.
[3] 坂巻顕太郎, 寒水孝司, & 濱崎俊光. (2019). 多重比較法. 朝倉書店.
[4] Dmitrienko, A., Tamhane, A. C., & Bretz, F. (Eds.). (2009). Multiple testing problems in pharmaceutical statistics. CRC press.
[5] Bretz, F., Pinheiro, J. C., & Branson, M. (2005). Combining multiple comparisons and modeling techniques in dose‐response studies. Biometrics, 61(3), 738-748.
[6] CHMP. (2017). Guideline on multiplicity issues in clinical trials. Available online at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/03/WC500224998.pdf
[7] FDA. (2017). Multiple endpoints in clinical trials: draft guidance for industry. Available online at: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM536750.pdf
[8] FDA. (2019). Enrichment strategies for clinical trials to support approval of human drugs and biological products: guidance for industry. Available online at: https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm332181.pdf
[9] Johnston, S., Pippen, J. Jr, Pivot, X., Lichinitser, M., Sadeghi, S., Dieras, V., Gomez, H. L., Romieu, G., Manikhas, A., Kennedy, M. J., Press, M. F., Maltzman, J., Florance, A., O'Rourke, L., Oliva, C., Stein, S., & Pegram, M. (2009). Lapatinib combined with letrozole versus letrozole and placebo as first-line therapy for postmenopausal hormone receptor-positive metastatic breast cancer. Journal of Clinical Oncology, 27(33), 5538-46.
[10] Lan, K. K. G., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. Biometrika, 70(3), 659-663.
[11] O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. Biometrics, 35(3), 549-556.
[12] Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. Biometrika, 64(2), 191-199.

3. 各規制当局の動向

この章では ICH および，各極の規制当局（FDA，EMA）が多重性の問題について示した指針を要約する。本文中では，多重評価項目に関する用語として“co-primary endpoint”，“multi-primary endpoint”，“主要副次評価項目”，“key secondary endpoint”，“複合評価項目”，“multi-component endpoint”，“composite endpoint”などの名称を用いたが，各規制当局間でこれら用語の定義についてコンセンサスが取られているわけではないため，試験実施計画書や統計解析計画書などにおける用語の使用には注意が必要である。

日米欧の 3 極は臨床試験のための統計的原則について協議し，合意された内容を 1998 年 2 月に ICH E9 ガイドラインとして最終化した（ICH, 1998）。ICH（1998）では，臨床試験の多重性に関する問題についても触れられている。日本ではガイドライン「臨床試験のための統計的原則」として 1998 年 11 月に通知ならびに施行された。その後，現在に至るまで日本の規制当局は多重性に関する公的な見解を示していないため，多重性に関しては日本における唯一の行政指針となっている。

欧州では 2002 年に医薬品委員会（EMA の委員会）が POINTS TO CONSIDER ON MULTIPLICITY ISSUES IN CLINICAL TRIALS を公開した（CPMP, 2002）。これは欧州における承認申請において重要と考えられた多重性の問題を記載した留意事項である。その後，2017 年 4 月に新たな規制上のアドバイス，推定における多重性，用量探索の新たなアプローチなどを加えたものとして，Guideline on multiplicity issues in clinical trials と題したガイドラインのドラフトを公開した（CHMP, 2017）。

米国では臨床試験の多重評価項目の問題と結果の解釈およびその問題をどのように取り扱うべきかに関する FDA の考えを示すものとして，Multiple Endpoints in Clinical Trials Guidance for Industry と題したガイダンスが作成中であり，そのドラフトが 2017 年 1 月に公開された（FDA, 2017）。

これらの資料に示された各規制当局の多重性に対する指針の概要を 3.1～3.4 に示し，共通点および相違点について 3.5 に要約する。

3.1. ICH E9（1998）

- ✓ 主要評価項目の個数は出来るだけ 1 つにすること。
- ✓ 複数の主要評価項目を用いる場合には第 1 種の過誤を制御する方法を治験実施計画書に記載すること。
- ✓ 複数の主要評価項目を用いる場合に第 2 種の過誤が増大すること，それによって必要な被験者数が増大することについて注意を喚起している。
- ✓ 多重性を回避あるいは減じることが推奨される。
- ✓ 多重性が存在する場合，調整は常に考慮すべきであり，調整方法の詳細，又はなぜ調整は必要ないと考えるのかという説明は，統計解析計画書に述べるべきである。
- ✓ 安全性および忍容性の変数に対して検定を用いる場合，通常は第 1 種の過誤よりも第 2 種の過誤により注意を払うべきである。
- ✓ 副次評価項目について，その数を試験で答えるべき限られた少数の問題と関連して制限すべきだとしているが，副次評価項目が複数あることに由来する多重性の問題については述べられていない。
- ✓ 指定した全ての主要変数において有効性を示すことが試験の目的（co-primary endpoint）である場合，第 1 種の過誤を調整する必要はないが，第 2 種の過誤および

必要な被験者数への影響は慎重に考慮すべき。

3.2. CPMP（2002）

- ✓ 治験実施計画書や統計解析計画書に多重性の調整方法を事前記載することを推奨。
- ✓ 複数の主要評価項目を用いる場合に第 2 種の過誤が増大すること，それによって必要な被験者数が増大することについて注意を喚起。
- ✓ 複数個の主要評価項目が必要な場合があることを認め，その場合，"closed testing procedure"に属する方法を用いれば有意水準の調整が不要（資料作成者注：予め仮説の順序を規定して検定を行う方法を用いれば有意水準の調整が不要になる場合がある）。
- ✓ 多重性を調整する方法を選択する際に検定結果と整合する信頼区間を構成できるかどうか，結果について臨床的な解釈ができるかどうかを考慮する必要がある。
- ✓ 一般的でない方法を使う場合には規制当局との相談を推奨。
- ✓ 安全性の変数であってもその検定結果が承認や labeling claim に反映される場合には主要変数と同様に多重性の問題を扱うべき。
- ✓ 臨床推奨用量の決定を目的とした用量反応試験に対しては FWER の制御が必要。
- ✓ 主要評価項目が達成された場合に副次評価項目の検証が可能。
- ✓ 副次評価項目が claim を意図しない場合には信頼区間や検定結果は探索的位置付け。
- ✓ composite endpoint を「精神神経領域などで使用される rating scale」と，「生存時間解析に関連するもの」の 2 種類に大別。
- ✓ 主要評価項目としての composite endpoint を用いる場合には，臨床的に重要な構成要素について解析することを推奨。
  - ➢ 構成要素の一部に基づいて claim を行うならば，事前の特定と適切な検証的統計解析計画が必要。

3.3. CHMP（2017）

- ✓ 治験実施計画書や統計解析計画書に多重性の調整方法を事前記載することを推奨。
- ✓ 複数の主要評価項目を用いる場合に第 2 種の過誤が増大すること，それによって必要な被験者数が増大することについて注意を喚起。
- ✓ 多重性を調整する方法を選択する際に検定結果と整合する信頼区間を構成できるかどうか，結果について臨床的な解釈ができるかどうかを考慮する必要あり。
- ✓ 一般的ではない方法を使う場合には規制当局と相談することを推奨。
- ✓ 安全性の変数であっても，検定結果が承認や labeling claim に反映される場合には主要変数と同様に多重性の問題を扱うべき。
- ✓ 用量反応試験では検定よりも推定にフォーカスすべき。
  - ➢ 統計学的に有意な最低の用量を見出すという手順はあまり価値がなく結果の解釈を誤ることがある。
  - ➢ 第 II 相試験では検証的な試験でない限り，FWER を制御するための群間比較の多重性の調整は不要 [8]。
- ✓ 主要評価項目が達成された場合に副次評価項目の検証が可能。
- ✓ 副次評価項目が claim を意図しない場合には信頼区間や検定結果は探索的位置付け。
- ✓ 特定のサブグループについて信頼性のある結論を得るためには，事前宣言と適切な解析計画が必要。

[8] EMA では用量反応関係の解析における MCP-Mod の利用について Qualification paper を作成（2014）

- ✓ composite endpoint を「精神神経領域などで使用される rating scale」と，「生存時間解析に関連するもの」の 2 種類に大別。
- ✓ 主要評価項目としての composite endpoint を用いる場合には，臨床的に重要な構成要素について解析することを推奨。
  - ➢ 構成要素の一部に基づいて claim を行うならば，事前の特定と適切な検証的統計解析計画が必要。
- ✓ 検定における多重性のみではなく，推定における多重性の問題について言及。

3.4. FDA（2017）

- ✓ 治験実施計画書や統計解析計画書に多重性の調整方法の事前記載を推奨。
- ✓ 複数の主要評価項目を用いる場合に第 2 種の過誤が増大すること，それによって必要な被験者数が増大することについて注意喚起。
- ✓ 開発の早期段階から薬剤効果を広く検討しておくことは，複数の候補となる評価項目からの感度が良い有益な評価項目の選択に寄与し，検証的試験では単一の主要評価項目を用いることにつながりうる。
- ✓ 安全性の変数であってもその検定結果が承認や labeling claim に反映される場合には主要変数と同様に多重性の問題を扱うべき。
- ✓ 主要評価項目で有効性が示された場合に限定して副次評価項目の検証が可能。
  - ➢ 検証を意図しない評価項目は副次評価項目ではなく探索的評価項目
  - ➢ 全ての主要評価項目と全ての副次評価項目を含む仮説族に対して FWER の強い制御が必要。
  - ➢ 検出力の低下を避けるために副次評価項目の個数を限定することを推奨。
- ✓ 複数の要素から成り立っている評価項目を multi-component endpoint と定義し，その中でもイベント発現までの時間を解析対象とする場合を composite endpoint としている。
- ✓ 薬剤が composite endpoint の全ての構成要素に対して好ましい効果を持つ確証はないことから，構成要素は個別に解析し常に試験の報告書に含めるべき。
  - ➢ 幾つかの構成要素について薬剤効果を示す場合は，多重性を考慮して予め統計解析計画書に仮説を記載すべき。
  - ➢ 構成要素の解析をするにあたり，各被験者について「最初に生じたイベントのみを考慮するアプローチ」と「全ての種類のイベントを考慮するアプローチ」の 2 種類が存在し，それぞれに特徴がある事を例示。
  - ➢ FDA ガイダンス“Clinical Studies Section of Labeling for Human Prescription Drug and Biological Products – Content and Format”では，構成要素が事前に独立した評価項目と規定され，事前に定義された仮説と統計解析計画により多重性を調整した下で有意とならない限り，labeling には構成要素の統計解析結果を示さないことを求めている。
    - ✧ 通常この場合の解析計画は「全ての種類のイベントを考慮するアプローチ」に基づく。
- ✓ 多重性に対する戦略と調整方法は数多く存在し，それぞれ利点と欠点がある。
  - ➢ 適切な戦略と調整方法の選択は，試験の計画段階で解決すべき課題であり，統計家の参画が推奨される。

3.5. 各指針の共通点および相違点

ICH E9（1998），CPMP（2002），CHMP（2017），FDA（2017）とも，有効性・安全性の評価に対しての多重性の問題に関する指針に大きな違いは認められないが，主な相違点

としてはガイダンスの適用範囲，用語の定義などが挙げられる。なお，日本に関しては独自の行政指針は存在せず，添付文書の記載方法も疾患単位である一方，欧米では多重性を適切に考慮した下での有意な結果は labeling claim に記載できる可能性があるなど，当局間での結果の解釈の差異も存在する。また，非劣性が達成された後に優越性を検証するデザイン（いわゆる，スイッチング）では，多重性の調整は FWER の適切な制御の枠組みに置いては理論上不要であるものの，規制当局間で結果の受け入れ可能性が異なることがある。国際共同試験では，承認申請後に結果の受け入れ可能性について当局間で差異が生じ，議論となることがあるため，試験開始前にこの点についても想定し，当局と十分に議論しておくことが望ましい。下記に主な共通点と相違点を記載する。なお Appendix 1 に共通点と相違点を整理した。必要に応じて参照されたい。

<u>共通点</u>

- ✓ 主要評価項目は出来るだけ 1 つにすべきだが，複数の主要評価項目を用いる際には多重性に注意
  - ➢ 検証試験では全体の第 1 種の過誤確率（FWER）の厳密な制御が必要
  - ➢ 複数の主要評価項目の<u>すべてが有意である時に試験成功とする場合（co-primary の場合）</u>は第 1 種の過誤確率の調整は不要だが，第 2 種の過誤確率は増大するため，症例数設計への影響を考慮
  - ➢ 複数の主要評価項目のうち，<u>いずれか 1 つが有意である時に試験成功とする場合</u>は，多重性の調整が必要
- ✓ 安全性評価項目であっても，効能・効果に含めたい場合は，有効性と同様に多重性の調整が必要
- ✓ 安全性評価では，安全性上のシグナルを見逃さない観点で第 1 種の過誤よりも第 2 種の過誤を重視すべき
- ✓ 検証したい評価項目およびその検定仮説，多重性の調整方法などは事前に試験実施計画書や統計解析計画書に規定されるべき

<u>相違点</u>

- ✓ ICH（1998），CPMP（2002），CHMP（2017）では，ガイダンスの適用範囲を多重性一般としているが，FDA（2017）では multiple endpoints に限定している。
- ✓ 用語の定義
  - ➢ CHMP（2017）では主要評価項目の他に検定を実施する評価項目として主要副次評価項目（"key secondary endpoint”）という用語を用いている。一方，FDA（2017）では主要副次評価項目にあたる用語は副次評価項目（"secondary endpoint”）とし，検定を実施しない評価項目を探索的評価項目（"exploratory endpoint”）としている。
  - ➢ CPMP（2002）, CHMP（2017）は，composite endpoint を 1）rating scale（特定の疾患で長年の使用経験があるもの）と，2）複数のイベント（死亡，心筋梗塞，障害を残すような脳卒中など）のいずれかが生じるまでの時間を用いるもの，の 2 つとしている。一方，FDA（2017）では，composite endpoint は CPMP（2002），CHMP（2017）における後者のみを指し，前者は multi-component endpoint として区別している。
- ✓ CPMP（2002）では，用量反応試験であっても第 1 種の過誤確率の厳密な制御を必須としていたが，CHMP（2017）では第 II 相試験においては必須ではないとしている。
- ✓ CHMP（2017）では，検定に伴う多重性のみではなく，推定における多重性についても言及している。

【参考文献】

[1] CPMP. (2002). Points to consider on multiplicity issues in clinical trials. Available online at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003640.pdf

[2] CHMP. (2013). Qualification opinion of MCP-Mod as an efficient statistical methodology for model-based design and analysis of Phase II dose finding studies under model uncertainty. Available online at: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2014/02/WC500161027.pdf

[3] FDA. (2006). Clinical Studies Section of Labeling for Human Prescription Drug and Biological Products — Content and Format: guidance for industry. Available online at: https://www.fda.gov/media/72140/download

[4] CHMP. (2017). Guideline on multiplicity issues in clinical trials. Available online at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/03/WC500224998.pdf

[5] FDA. (2017). Multiple endpoints in clinical trials: draft guidance for industry. Available online at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM536750.pdf

4. 多重性調整の事例

4.1. 審査において多重性の問題が議論となった事例

本章では，承認審査過程において，多重性の問題が議論となった事例や複雑な多重性調整の方法など興味深い事例を紹介する。欧米の規制当局の審査事例も含まれるが，国際共同開発が主流となりつつあることから，本邦の臨床開発担当者への示唆を含む事例と考えている。これらの事例を通じて，臨床試験の計画段階から統計担当者だけではなく関係者が協働して多重性の問題について検討することの必要性を強調したい。

4.1.1. Plegridy

当局から仮説構造の変更を指示された事例である。Plegridy（pegylated interferon beta-1a）は再発寛解型多発性硬化症の治療薬として 2014 年 7 月に EMA，8 月に FDA に承認されている（本邦では未開発）。EMA や FDA の申請資料から，主要評価項目に加え，主要副次評価項目を含めた複数の評価項目間の多重性調整を行っていることが分かる。これは統計的有意差かつ臨床的意義が認められた主要副次評価項目は，添付文書の効能・効果へ反映できることがあるためであろう。また，効能・効果へ反映できれば，既存薬との添付文書（欧州製品情報概要 SmPC，米国添付文書情報 labeling）での差別化を狙ったものかもしれない。

以下は，CHMP による Assessment Report において，主要副次評価項目の検定順序について議論された内容である。

<u>多重性に関連した審査上の論点</u>

検証試験である 105MS301 試験のデザイン概要を表 4.1.1-1 に示す。主要評価項目に加えて検証を意図した主要副次評価項目が重要度の高いものから 3 つ設定されている。投与群は 3 群設定されており，主要評価項目および各主要副次評価項目に対し Q2W 群 vs プラセボ群の検定を行い，統計学的有意差が得られた場合は各評価項目について同様に Q4W 群 vs プラセボ群の検定を行う仮説構造となっている。多重性は固定順序法[9]により調整した。

表 4.1.1-1　105MS301 試験のデザインの概要

Phase	III
試験デザイン	二重盲検並行群間比較国際共同試験
投与群	● プラセボ群 ● 125 μg 4週間隔投与群（Q4W群） ● 125 μg 2週間隔投与群（Q2W群）
主要評価項目	1 人年あたりの再発率
主要副次評価項目	● MRIによるT2新規・拡大病巣数 ● 1年時点での再発患者の割合 ● 身体機能障害の進行が認められた患者の割合

[Assessment Report, Table 12 より改変]

解析結果を表 4.1.1-3 に示す。CHMP は下記のとおり，申請者が主要副次評価項目の第 3 位に位置づけた「身体機能障害の進行が認められた患者の割合」を，審査上は最重要な主

[9] 注）「階層的検定手順法」や「Fixed sequence 法」と呼ばれることもある。また，慣例的に「閉（検定）手順」とも呼ばれることがある。

要副次評価項目として位置づけたことを指摘している（表 4.1.1-2）。

Disability (3-month sustained disability progression) was included as a secondary endpoint, however, it was only ranked 3rd among the secondary endpoints. In this assessment, disability was considered to be the most important secondary endpoint.

表 4.1.1-2　申請者と当局間での評価項目の検定順序の相違点

評価項目	申請者により示された 検定順序	当局が重視した 検定順序
主要	1 人年あたりの再発率	
主要副次	① MRI による T2 新規・拡大病巣数 ② 1 年時点での再発患者の割合 ③ <u>身体機能障害の進行が認められた患者の割合</u>	① <u>身体機能障害の進行が認められた患者の割合</u> ② MRI による T2 新規・拡大病巣数 ③ 1 年時点での再発患者の割合

これは，CHMP の多発性硬化症の疾患評価ガイドライン（EMA, 2015）において，下記の言及があることによるものと考えられる。

A relapse-based primary endpoint though cannot be taken as a surrogate for disability progression and this would be expressed accordingly in the SmPC. Moreover, if the primary endpoint is based on relapse assessment, progression of disability should be evaluated as key secondary endpoint.

固定順序法は，予め設定された順序に従って検定が行われ，統計学的有意差が認められなくなった時点で，下位の順序に設定された有効性評価項目の検定を実施しないという多重性調整法である。本試験では，結果的に主要評価項目および全ての主要副次評価項目で統計学的有意差が認められたため，申請者と当局の考える検定順序の相違は大きな問題とならなかった。

しかしながら，仮に，当局が最重要な主要副次評価項目として指摘した評価項目（申請者が第 3 位に位置づけた主要副次評価項目）において，統計学的有意差が認められなかった場合を考えてみる。この場合，申請者が第 1 位および第 2 位に位置づけた主要副次評価項目が統計学的に有意であったとしても，多重性を調整した上で有意となった結果は主要評価項目のみとなるため，添付文書（欧州製品情報概要 SmPC）への主要副次評価項目結果の記載は認められなかったかもしれない。

本事例は，多重性の問題が統計学的な観点だけで解決できるものではなく，疾患評価ガイドラインなどを踏まえ，臨床的観点からの検討も必要になりうることを示している。本事例では，承認申請後に申請者の考える検定順序と当局の考える検定順序の差異について議論されたものであり，仮説構造については予め規制当局と合意しておくことが望ましいことが分かる。

表 4.1.1-3　105MS301 試験の解析結果

評価項目		Q4W vs プラセボ	Q2W vs プラセボ
主要評価項目 1 人年あたりの再発率	プラセボに対する減少率	27.5 %	35.6 %
	p 値	0.0114	0.0007
主要副次評価項目 1) MRI による T2 新規・拡大病巣数	プラセボに対する減少率	28 %	67 %
	p 値	0.0008	0.0001
主要副次評価項目 2) 1 年時点での再発患者の割合	プラセボに対する減少率	26 %	39 %
	p 値	0.0200	0.0003
主要副次評価項目 3) 身体機能障害の進行が認められた患者の割合	プラセボに対する減少率	38 %	38 %
	p 値	0.0380	0.0383

[Assessment Report, Table 12 より改変]

【参考文献】

[1] Plegridy Assessment Report. Available online at: http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/002827/WC500170303.pdf

[2] EMA. (2015). Guideline on clinical investigation of medicinal products for the treatment of multiple sclerosis. Available online at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2015/03/WC500185161.pdf

4.1.2. Spiolto Respimat

規制当局間で異なる仮説構造を提示された事例である。Spiolto Respimat（チオトロピウム臭化物／オロダテロール塩酸塩）は慢性肺閉塞性疾患（COPD）の治療薬として，2015 年 5 月に欧米で，2015 年 8 月に本邦で承認されている。

多重性に関連した審査上の論点

日米欧の三極での申請を意図した 2 つの検証的国際共同治験（1237.5 試験および 1237.6 試験）において，EMA と FDA では，有効性の主要評価に関する疾患評価ガイドライン上の要件が異なったため，試験実施計画書において事前に EMA と FDA で固定順序法による検定の仮説構造をそれぞれ別個に定めた旨が CTD に記載されている。具体的には，EMA 申請用の仮説構造では，CHMP の COPD 疾患評価ガイドライン（EMA, 2012）に従って，主要評価項目に疾患特異的 QOL スコアである SGRQ 総スコア，同様に主要副次評価項目として症状に対する TDI 総スコアが含まれている。一方，FDA 申請用の仮説構造では，呼吸機能に関する評価項目のみが仮説構造に含まれている。なお，本邦の申請においては FDA 申請用と同様の仮説構造を用いている。

仮説構造を下表に示す。投与群および評価項目が複数存在することから，複雑な仮説構造となっている。投与群は，本剤はチオトロピウムとオロダテロールの配合剤であるため，各評価項目に対して，配合の妥当性検証のために，チオトロピウム単剤およびオロダテロール単剤の各々に対する優越性の検証が行われている。また，チオトロピウムは，配合剤，単剤ともに 2 用量（2.5 μg / 5 μg）が設定されている。なお，SGRQ 総スコアおよび TDI 総スコアは，1237.5 試験および 1237.6 試験を併合したうえで評価を行っている。

表 4.1.2-1　1237.5 試験および 1237.6 試験の仮説構造

評価項目	被験薬群	対照薬群	FDA 用仮説構造での順位	EMA 用仮説構造での順位
主要評価項目				
Mean FEV_1 $AUC_{0\text{-}3h}$	Tio + Olo 5/5 µg	Olo 5 µg	1	1
		Tio 5 µg	2	2
Mean トラフ FEV_1	Tio + Olo 5/5 µg	Olo 5 µg	3	3
		Tio 5 µg	4	4
Mean SGRQ 総スコア（2 試験併合）	Tio + Olo 5/5 µg	Olo 5 µg	-	5
		Tio 5 µg	-	6
Mean FEV_1 $AUC_{0\text{-}3h}$	Tio + Olo 2.5/5 µg	Olo 5 µg	5	7
		Tio 2.5 µg	6	8
Mean トラフ FEV_1	Tio + Olo 2.5/5 µg	Olo 5 µg	7	9
		Tio 2.5 µg	8	10
Mean SGRQ 総スコア（2 試験併合）	Tio + Olo 2.5/5 µg	Olo 5 µg	-	11
		Tio 2.5 µg	-	12
主要副次評価項目				
TDI 総スコア（2 試験併合）	Tio + Olo 5/5 µg	Olo 5 µg	-	13
		Tio 5 µg	-	14
	Tio + Olo 2.5/5 µg	Olo 5 µg	-	15
		Tio 2.5 µg	-	16
副次評価項目				
Mean FEV_1 $AUC_{0\text{-}3h}$	Tio + Olo 2.5/5 µg	Tio 5 µg	9	17
Mean トラフ FEV_1		Tio 5 µg	10	18
Mean SGRQ 総スコア（2 試験併合）		Tio 5 µg	-	19
TDI 総スコア（2 試験併合）		Tio 5 µg	-	20

Olo：オロダテロール，Tio：チオトロピウム
Tio + Olo：チオトロピウム＋オロダテロール配合剤
[チオトロピウム臭化物水和物・オロダテロール塩酸塩 CTD 2.7.3B，表 1.4.2: 1 より改変]

【参考文献】

[1] チオトロピウム臭化物水和物・オロダテロール塩酸塩 CTD. Available online at: http://www.pmda.go.jp/drugs/2015/P20150908001/index.html

[2] EMA. (2012). Guideline on clinical investigation of medicinal products in the treatment of chronic obstructive pulmonary disease (COPD). Available online at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/08/WC500130880.pdf

4.1.3. Yervoy

ヤーボイ（イピリムマブ（遺伝子組換え））は根治切除不能な悪性黒色腫の治療薬として，2015 年 7 月に本邦で承認されている（効能効果の表現は完全には一致しないが，米国では 2011 年 3 月に，欧州では 2011 年 7 月に悪性黒色腫の治療薬として承認されている）。

以下は，PMDA の承認審査過程において多重性の調整が行われていないことが議論された事例である。

<u>多重性に関連した審査上の論点</u>

海外第 III 相試験（MDX010-20 試験）は，前治療歴を有する根治切除不能な悪性黒色腫患者を対象とし，無作為化された 676 例（本薬/gp100 併用投与群（以下「本薬/gp100 併用群」）403 例，本薬/プラセボ併用投与群（以下「本薬群」）137 例，プラセボ/gp100 併用投与群（以下「gp100 群」）136 例）が，ITT 集団として，有効性の解析対象とされた。

主解析とされた，本薬/gp100 併用群と gp100 群との全生存期間（OS）の比較において，本薬/gp100 併用群の優越性が検証された（表 4.1.3-1）。

表 4.1.3-1 主解析の結果（ITT 集団）

	本薬/gp100 併用群	gp100 群
例数	403	136
死亡数（%）	306（76.0）	119（87.5）
中央値［95% CI］（カ月）	9.95［8.48，11.50］	6.44［5.49，8.71］
ハザード比［95% CI］	0.68［0.55，0.85］	
P 値（両側）	0.0004	

副次解析とされた OS の結果を表 4.1.3-2，表 4.1.3-3 に示す。申請者は，本薬群でも gp100 群を上回る傾向が認められ，かつ本薬群と本薬/gp100 併用群で明確な違いは認められなかったことから，本薬の単独投与により OS の延長が期待できると考えた。ただし，本薬群と gp100 群の比較において多重性の調整は行われなかった。

表 4.1.3-2 副次解析の結果（本薬剤群と gp100 群の比較，ITT 集団）

	本薬群	gp100 群
例数	137	136
死亡数（%）	100（73.0）	119（87.5）
中央値［95% CI］（カ月）	10.12［8.02，13.80］	6.44［5.49，8.71］
ハザード比［95% CI］	0.66［0.51，0.87］	
P 値（両側）	0.0026	

表 4.1.3-3 副次解析の結果（本薬/gp100 併用群と本薬群の比較，ITT 集団）

	本薬/gp100 併用群	本薬群
例数	403	137
死亡数（%）	306（75.9）	100（73.0）
中央値［95% CI］（カ月）	9.95［8.48，11.50］	10.12［8.02，13.80］
ハザード比［95% CI］	1.04［0.83，1.30］	
P 値（両側）	0.7575	

しかし，PMDA は，前治療歴を有する根治切除不能な悪性黒色腫患者に対して gp100 は未承認であり，臨床的位置付けなどが明らかではないことなどから，当該試験成績を基に，今般の承認申請における申請用法・用量である本薬単独投与の有効性について評価することには限界があると考えた。また，本薬群と gp100 群の比較については多重性の調整が行われていないことなどから，MDX010-20 試験成績を基に，gp100 群に対する本薬群の OS の優越性が検証されたとは判断できないと考えた。

以上のような問題点があるものの，対照群として gp100 が設定された経緯，MDX010-20 試験は OS を主要評価項目とした無作為化比較試験であることなどを考慮し，根治切除不能な悪性黒色腫患者の生存に及ぼす gp100 の影響，MDX010-20 試験の各群における OS の成績なども確認した上で，MDX010-20 試験成績を中心に，本薬単独投与の有効性について総合的に評価することとされた。

PMDA は，下記の点を考慮し，前治療歴を有する根治切除不能な悪性黒色腫患者に対する本薬単独投与の一定の有効性は示されたと判断した。

- ✓ gp100 群に対する本薬/gp100 併用群の OS の優越性が示されたこと。
- ✓ gp100 群と比較して本薬群で OS が延長する傾向が認められたこと。
- ✓ 本薬/gp100 併用群と本薬群の OS に明確な差異は認められていないこと。
- ✓ 申請者が実施した過去の臨床試験成績に基づく考察などを考慮すると，gp100 投与により，根治切除不能な悪性黒色腫患者の OS に悪影響を及ぼす可能性は低いと考えられること。

【参考文献】

[1] イピリムマブ（遺伝子組換え）審議結果報告書・審査報告書（2015 年 07 月 03 日）. Available online at: http://www.pmda.go.jp/drugs/2015/P20150722002/67060500022700AMX00696000_A100_1.pdf

[2] イピリムマブ（遺伝子組換え）CTD. Available online at https://www.pmda.go.jp/drugs/2015/P20150722002/index.html

4.2. 複雑な多重性の調整方法

4.2.1. Olumiant

オルミエント錠（一般名：バリシチニブ）は，既存治療で効果不十分な関節リウマチの治療薬として，2017 年 5 月に本邦および欧州で，2018 年 6 月に米国（2 mg のみ）で承認されている。本品目では，4 本の第 III 相試験（I4V-MC-JADW 試験，I4V-MC-JADX 試験，I4V-MC-JADV 試験，I4V-MC-JADZ 試験：以降は共通である「I4V-MC-」の部分を省略する）で，Plegridy（4.1.1）や Spiolto Respimat（4.1.2）と同様に複数の評価項目間並びに複数の投与群間の多重性調整が行われたが，その仮説構造が特徴的であるため紹介する。

多重性に関する論点

表 4.2.1-1 に試験ごとの評価項目の優先順位を示す。JADW および JADX 試験では，ゲートキーピング法が用いられているため，固定順序法と同様に基本的に優先順位が上位の評価項目で有意差が認められた場合のみ，下位の評価項目の検定が可能となる。同一の優先順位となっている評価項目は，Hochberg の方法を用いて評価項目 A と評価項目 B がα=0.05 で両方有意の場合，さらに下位の検定が可能となる。

JADV および JADZ 試験は，上記 2 試験と同様に評価項目間の優先順位をつけて順番に検定をしていくが，より複雑な構造を定義しており，そのような場合でも仮説構造を視覚的に理解しやすいグラフィカル・アプローチという方法が用いられた。詳細は，図 4.2.1-1 および図 4.2.1-2 で説明する。

主要評価項目である ACR20 改善基準の優先順位が最上位であることは 4 試験で共通となっているが，副次評価項目の優先順位は試験間で異なる設定がなされていることが分かる。

表 4.2.1-1　4 試験の投与群並びに代表的な評価項目の優先順位の設定

試験	JADW	JADX	JADV	JADZ
投与群	4 mg 群 2 mg 群 プラセボ群	4 mg 群 2 mg 群 プラセボ群	4 mg 群 対照薬群 プラセボ群	4 mg 単独群 MTX 単独群 併用群
主要評価項目				
ACR20 改善基準	1	1	1	1
副次評価項目				
HAQ-DI	2	2	2	3
DAS28-hsCRP	2	2	3	2
SDAI 寛解率	3	3	4	5
mTSS	－	－	2	4
朝のこわばりの持続時間	3	3	5	X
朝のこわばりの重症度	－	4	6	X
最もひどい疲労	X	4	7	X
最も強い疼痛	X	4	7	X

表中の数値：各試験内における評価項目の検定順序
X：多重性の調整を行っていない項目
－：評価項目として設定されていない項目

JADW 試験および JADX 試験の仮説構造

表 4.2.1-1 に示したとおり，両試験における有効性評価項目の優先順位は類似している。しかしながら，投与群間の比較を加味した場合，JADW 試験では，有効性評価項目より用量間の比較を優先しているのに対し，JADX 試験では有効性評価項目と用量間の比較のバランスを取りながら検定を実施していることが分かる。

両試験ともに，多重性を調整したすべての評価項目において有意差が認められた。

表 4.2.1-2　JADW および JADX 試験の仮説構造

試験	JADW		JADX	
投与群の比較	4 mg vs プラセボ	2 mg vs プラセボ	4 mg vs プラセボ	2 mg vs プラセボ
ACR20	1	4	1	3b-1
HAQ-DI	2	5	2	3b-2
DAS28-hsCRP	2	5	2	3b-2
SDAI 寛解率	3	6	3a	4
こわばり持続時間			3a	－

こわばり重症度			4	－
最もひどい疲労			4	－
最も強い疼痛			4	－

JADV 試験および JADZ 試験の仮説構造

両試験ともに複雑な仮説構造を定義しているため，グラフィカル・アプローチという方法を用いて表現されている。仮説は○（ノードと呼ぶ）で表されており，当該の仮説が棄却された（有意差が認められた）場合に，ノードから出ている矢印に記載された係数（係数の合計は 1 になる）を乗じた有意水準αが，矢印の先のノードの仮説に分配される。

JADV 試験（図 4.2.1-1）では，H1 の仮説検定を有意水準両側 5%（α=0.05）で実施し，棄却された場合，H2 と H3 のそれぞれにα=0.05×0.5=0.025 が分配される。

本仮説構造では，例えば H8 から H2 のように矢印が循環している構造が見て取れる。この構造の下では当該仮説が棄却されなかったとしても他の仮説が棄却された場合に係数に応じた有意水準を“再利用”することが可能なため再度，仮説検定を実施することができる。

JADZ 試験（図 4.2.1-2）では，主要目的である H_{M0} を含む H_{M-}の仮説群が被験薬 4 mg 単独群と MTX 単独群との比較（図の左側），H_{C-}の仮説群が被験薬 4 mg＋MTX 併用群と MTX 単独群との比較（図の右側）で構成されている（H_{C1}～H_{C5} を H_{C-}，H_{M1} および H_{C1} を H_{-1} と表記する）。

最初に H_{M0} の仮説検定を有意水準両側 5%（α=0.05）で実施し，有意差が認められた場合，H_{M1} にα=0.05×0.1=0.005 が，H_{C1} にα=0.05×0.9=0.045 が分配される。これは，H_{C-}の仮説群，すなわち被験薬 4 mg＋MTX 併用群と MTX 単独群との比較において有意差が検出されやすい構造になっていることが分かる。しかしながら，図の下方にある検定（H_{-3} および H_{-4}）で有意差が認められた場合，他方の投与群の比較に用いることができるような手当がなされている（H_{M3}, H_{M4}→H_{C1} および H_{C3}, H_{C4}→H_{M1}）。

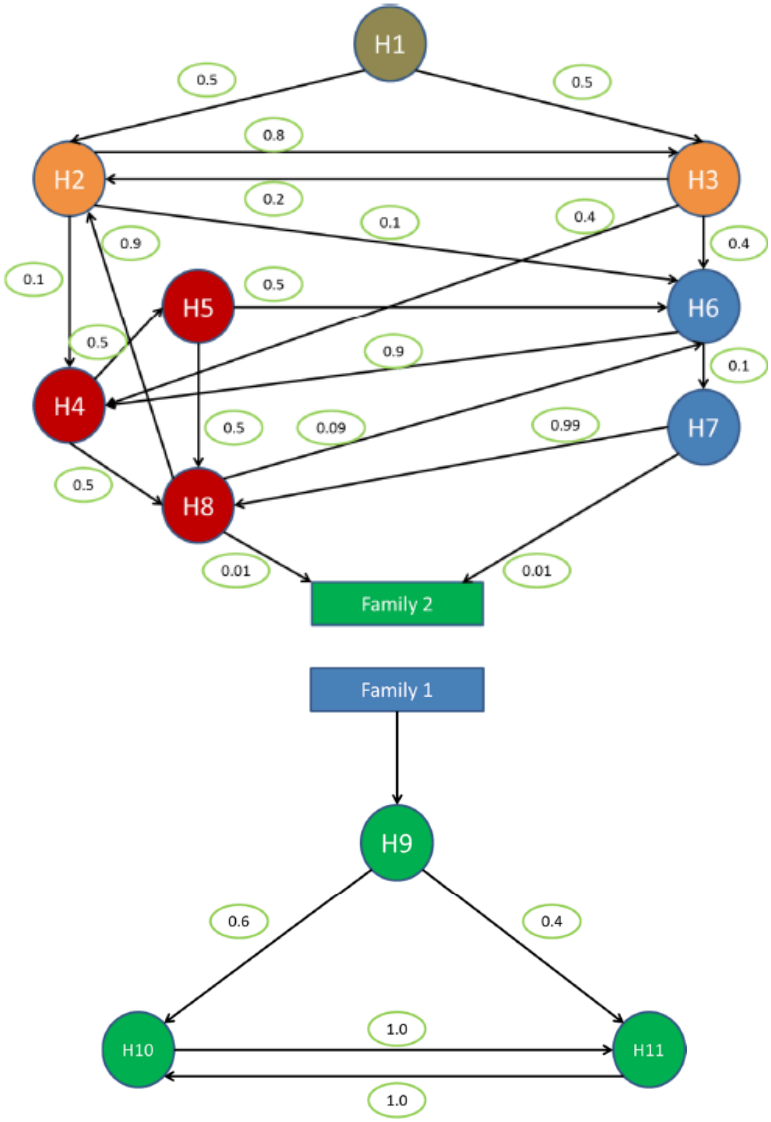
結論として，JADV 試験は，多重性を調整したすべての評価項目において，有意差が認められた（結果は割愛）。

JADZ 試験は，H_{M4} のみ有意差が認められず，それ以外の帰無仮説では有意差が認められた（表 4.2.1-3）。

表 4.2.1-3　JADZ 試験の検定結果

H_{M-}：4 mg vs MTX		H_{C-}：4 mg＋MTX vs MTX	
H_{M0}	P<0.001*	－	－
H_{M1}	P=0.003*	H_{C1}	P<0.001*
H_{M2}	P<0.001*	H_{C2}	P<0.001*
H_{M3}	P<0.001*	H_{C3}	P<0.001*
H_{M4}	P=0.158	H_{C4}	P=0.026*
H_{M5}	P=0.003*	H_{C5}	P<0.001*

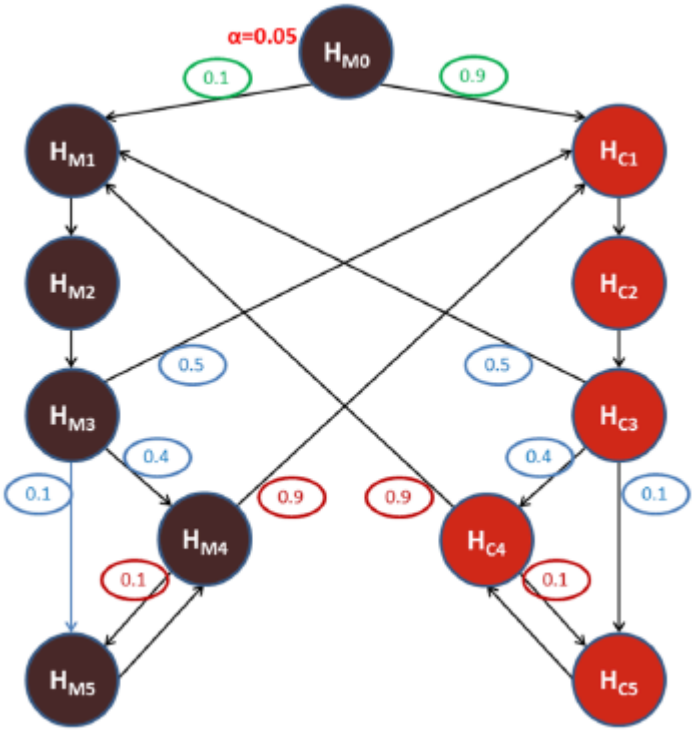
*：多重性調整後でも有意となった仮説



[オルミエント錠 4 mg オルミエント錠 2 mg CTD 2.7.6, 図 2.7.6.3.6-2 より改変]

図 4.2.1-1 JADV 試験の仮説構造

H1：ACR20，H2：mTSS，H3：HAQ-DI，H4：DAS28-hsCRP，H5：SDAI 寛解率，H6：ACR20（非劣性），H7：DAS28-hsCRP，H8：こわばり持続時間，H9：こわばり重症度，H10：最もひどい疲労，H11：最も強い疼痛。H6 およびH7 は被験薬 4 mg群と対照薬群との比較，その他の仮説は被験薬 4 mg群とプラセボ群との比較。H6 のみ非劣性で，その他の仮説は優越性。Family 2 は仮説H9~H11 の仮説族を示し，同様にFamily 1 は仮説H1~H8 の仮説族を示す。H7 あるいはH8 が有意であればFamily 2 の検定に進むことができる。



[オルミエント錠 4 mg オルミエント錠 2 mg CTD 2.7.6, 図 2.7.6.3.7-2 より改変]

図 4.2.1-2 JADZ 試験の仮説構造

$H_{M\text{-}}$は被験薬 4 mg 単独群と MTX 単独群との比較，$H_{C\text{-}}$は被験薬 4 mg＋MTX 併用群と MTX 単独群との比較を示す。$H_{\text{-}0}$：ACR20，$H_{\text{-}1}$：ACR20，$H_{\text{-}2}$：DAS28-hsCRP，$H_{\text{-}3}$：HAQ-DI，$H_{\text{-}4}$：mTSS，$H_{\text{-}5}$：SDAI 寛解率。H_{M0}のみ非劣性仮説で，その他の仮説は優越性仮説。

【参考文献】

[1] バリシチニブ 審議結果報告書・審査報告書（2015 年 07 月 03 日）. Available online at: https://www.pmda.go.jp/drugs/2017/P20170724002/530471000_22900AMX00582_A100_1.pdf

[2] バリシチニブ CTD. Available online at: https://www.pmda.go.jp/drugs/2017/P20170724002/index.html

4.2.2. Ongentys

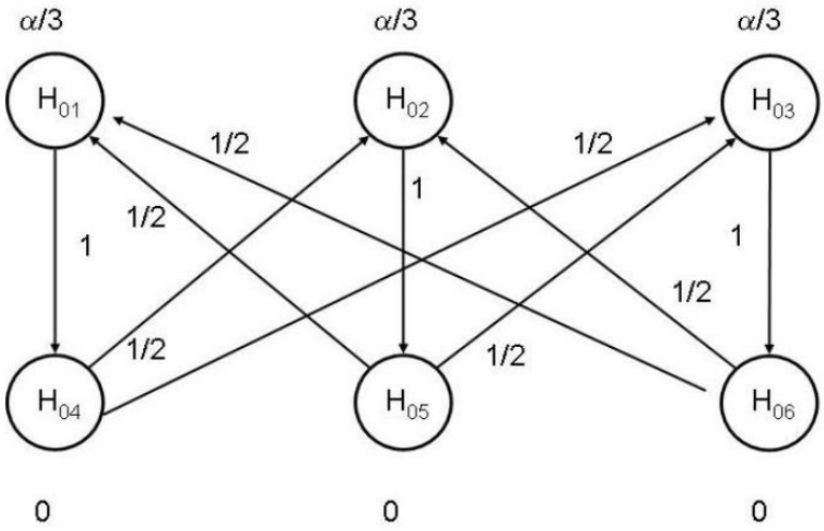
Ongentys（opicapone）は，レボドパ／ドパ脱炭素酵素阻害薬の併用療法で運動症状の日内変動が認められるパーキンソン病患者における補助療法として 2016 年 6 月に EMA で承認されている（本邦では 2019 年 2 月に承認申請を行っている）。

多重性に関する論点

第 III 相多施設二重盲検無作為化実薬およびプラセボ対照並行群間試験（BIA-91067-301）では，opicapone の 3 つの用量（5 mg，25 mg および 50 mg）のプラセボに対する優越性および実薬（entacapone 200 mg）に対する非劣性を評価した。二重盲検（DB）期で被験者は，opicapone 5 mg 群，opicapone 25 mg 群，opicapone 50 mg 群，entacapone 200 mg 群，プラセボ群のいずれかに 1:1:1:1:1 の比で無作為に割り付けられた。主要評価項目は，

オフ時間の絶対値のベースラインから DB 期終了時までの変化量であった。

BIA-91067-301 試験において，opicapone の各用量群とプラセボ群および entacapone 200 mg 群の比較によって生じる検定の多重性を調整するために用いられた検定手順を，グラフィカル・アプローチで表したものを図 4.2.2-1 に示した。



H_{01}, H_{02}, H_{03}：プラセボに対する opicapone の 5, 25, 50 mg の優越性仮説
H_{04}, H_{05}, H_{06}：entacapone 200 mg に対する opicapone の 5, 25, 50 mg の非劣性仮説
[Assessment Report, Figure3 を転載]

図 4.2.2-1　BIA-91067-301 試験の検定手順

同様にグラフィカル・アプローチで仮説構造を表現していた「4.2.1. Olumiant」では，起点となるノードが 1 つであったが，本事例では，起点となるノードが 3 つに分かれている点が特徴的である。図中の各ノードの上下に表示された数値は，検定手順の開始時に各ノードの帰無仮説に配分された有意水準である。図中の矢印の横に記載された数値は，矢印の出ている帰無仮説が棄却された（検定が有意であった）場合に，矢印の先の帰無仮説に配分される有意水準の重みを表している。重みが 1 の場合は，棄却された帰無仮説の有意水準が矢印の先の帰無仮説にそのまま分配され，重みが 1/2 の場合は，棄却された帰無仮説の有意水準が矢印の先の 2 つの帰無仮説に半分ずつ分配されることを意味する。具体的な検定手順は下記の通りである。

1. 検定仮説 H_{01}，H_{02}，H_{03} に対しα/3 の有意水準でそれぞれ検定する。
2. H_{04}，H_{05}，H_{06} のうち H_{01}，H_{02}，H_{03} の検定で棄却できた検定仮説の用量群については，同じ有意水準（α/3）で検定する。
3. 更に，H_{04}，H_{05}，H_{06} のうち棄却できた検定仮説の用量群について，H_{01}，H_{02}，H_{03} のうちまだ有意になっていない検定仮説の検定に対応する有意水準が再利用される。

BIA-91067-301 試験の主要評価項目の検定結果を表 4.2.2-1 に示した。opicapone 5 mg 群，25 mg 群および 50 mg 群のプラセボ群に対する優越性の検定の p 値は，0.059，0.107 および 0.001 であり，最初に配分された有意水準である 0.0167（=0.05/3）を下回って優越性が示されたのは，opicapone 50 mg 群のみであった。また，opicapone 50 mg 群の entacapone 200 mg 群に対する非劣性の検定の p 値は 0.003 であり，有意水準 0.0167 を下

回って非劣性も示された。そのため，opicapone 5 mg 群および 25 mg 群のプラセボ群に対する優越性を，有意水準 0.025（=0.0167+0.0167/2）で再度行ったが，どちらの用量群もプラセボ群に対する優越性を示すことが出来なかった。ここで検定手順が終了し，opicapone 5 mg 群および 25 mg 群については，entacapone 200 mg 群に対する非劣性の検定は行われていない。

表 4.2.2-1　BIA-91067-301 試験の主要評価項目の検定結果

検定仮説	P 値	有意水準
Test for superiority (FAS)		
opicapone 5 mg vs placebo（H_{01}）	0.059	0.025
opicapone 25 mg vs placebo（H_{02}）	0.107	0.025
opicapone 50 mg vs placebo（H_{03}）	0.001	0.0167
Test for non-inferiority (PP)		
opicapone 5 mg vs entacapone 200 mg（H_{04}）	0.080	N/A
opicapone 25 mg vs entacapone 200 mg（H_{05}）	0.073	N/A
opicapone 50 mg vs entacapone 200 mg（H_{06}）	0.003	0.0167

FAS: Full Analysis Set, PP: Per-protocol Set
[Assessment Report, Table17 より改変]

【参考文献】

[1] Ongentys Assessment Report. Available online at: https://www.ema.europa.eu/en/documents/assessment-report/ongentys-epar-public-assessment-report_en.pdf

4.3. 複合評価項目における事例

4.3.1. Brilinta

ブリリンタ錠（一般名：チカグレロル）は，「アテローム血栓症の発症リスクが特に高い陳旧性心筋梗塞」および「経皮的冠動脈形成術が適用される急性冠症候群」を適応として，2010 年 12 月に欧州，2011 年 7 月に米国，2016 年 9 月に本邦で承認されている。ここでは，PMDA による審査の過程において複合評価項目の構成要素に関する検討が行われた「経皮的冠動脈形成術が適用される急性冠症候群」に関する試験について紹介する。

多重性に関連した審査上の論点

海外で実施された急性冠症候群患者約 18,000 例を対象とした国際共同第 III 相試験（PLATO 試験）では，有効性の主要評価項目とされた心血管死，心筋梗塞（無症候性除く），脳卒中の複合評価項目の抑制効果について，国内外における標準薬であるクロピドグレルに対する本薬の優越性が検証され，海外各国で承認された。

日本における開発が海外から遅れて開始されたことから，当該試験に日本は参加できなかったため，経皮的冠動脈形成術が適用される急性冠症候群患者約 800 例を対象とし，PLATO 試験と同様の主要評価項目を設定したアジア共同第 III 相試験が実施された。その結果を表 4.4.1-1 に示す。

表 4.4.1-1　アジア共同第 III 相試験の全集団および日本人部分集団における有効性の各評価項目の発現状況（FAS）

評価項目／集団		AZD6140 群（401 例）		対照薬群（400 例）		ハザード比 [95%CI]
		発現例数（%）	KM%	発現例数（%）	KM%	
心血管死，心筋梗塞（無症候性除く），脳卒中	全集団	36 (9.0)	10.2	25 (6.3)	8.1	1.47 [0.88, 2.44]
	日本人	34 (9.4)	10.3	24 (6.7)	8.5	1.44 [0.85, 2.43]
心血管死，自然発症した心筋梗塞（無症候性除く），脳卒中	全集団	18 (4.5)	5.2	13 (3.3)	4.8	1.39 [0.68, 2.85]
	日本人	16 (4.4)	4.9	12 (3.3)	4.8	1.34 [0.63, 2.83]
死因を問わない死亡，心筋梗塞（無症候性除く），脳卒中	全集団	37 (9.2)	10.5	25 (6.3)	8.1	1.51 [0.91, 2.50]
	日本人	34 (9.4)	10.3	24 (6.7)	8.5	1.44 [0.85, 2.43]
心血管死，全ての心筋梗塞，脳卒中，重度の再発性心筋虚血，再発性心筋虚血，TIA，他の動脈性血栓イベント	全集団	38 (9.5)	10.8	32 (8.0)	10.3	1.20 [0.75, 1.93]
	日本人	36 (9.9)	10.9	31 (8.6)	10.9	1. 17 [0.73, 1.90]
心筋梗塞（無症候性除く）	全集団	24 (6.0)	7.1	15 (3.8)	4.5	1.63 [0.85, 3.11]
	日本人	22 (6.1)	6.9	14 (3.9)	4.6	1.60 [0.82, 3.12]
心血管死	全集団	9 (2.2)	2.5	7 (1.8)	1.6	1.28 [0.48, 3.45]
	日本人	9 (2.5)	2.7	6 (1.7)	1.5	1.49 [0.53, 4.20]
脳卒中	全集団	9 (2.2)	2.5	6 (1.5)	2.7	1.50 [0.54, 4.23]
	日本人	9 (2.5)	2.7	6 (1.7)	2.8	1.50 [0.53, 4.22]
死因を問わない死亡	全集団	10 (2.5)	2.8	7 (1.8)	1.6	1.42 [0.54, 3.74]
	日本人	9 (2.5)	2.7	6 (1.7)	1.5	1.49 [0.53, 4.20]

AZD6140 群：チカグレロル群，対照薬群：クロピドグレル群
KM%：Kaplan-Meier 法により算出した 12 ヶ月時点でのイベント発現率
[ブリリンタ錠 60 mg，ブリリンタ錠 90 mg 審査報告書 表 37 より改変]

有効性の主要評価項目とされた投与開始 12 か月時点の心血管死，心筋梗塞（無症候性除く），脳卒中の複合評価項目の KM%は，チカグレロル群で 10.2%，対照薬群で 8.1%であった。チカグレロル群の対照薬群に対するハザード比は 1.47 となり，PLATO 試験の結果（KM%：チカグレロル群で 9.8%，対照薬群で 11.7%，ハザード比：0.84）と異なりハザード比は 1 を上回った。また，複合評価項目の構成要素をはじめとした，いずれの副次評価項目においても，対照薬群と比較してチカグレロル群のイベント発現例数が多く，ハザード比は 1 を上回った。

その原因について，アジア共同第 III 相試験が有効性の検証を目的としない症例数で実施されたことの他に，依頼者は，手技に関連した心筋梗塞を除いた複合イベント（心血管死，自然発症した心筋梗塞（無症候性除く），脳卒中）についての事後解析にて，イベント発現数ならびに KM%で顕著な群間差が認められなかったことや，PLATO 試験の国別サブグループの主要評価項目に関するハザード比の分布範囲と比較して上限付近に位置するものの，分布の範囲内であったことなどを説明したが，PMDA はいずれも推測の域を出ず，原因は不明であると結論づけた。その上で，PLATO 試験の成績や類薬の承認状況を踏まえ，一定の有効性を示す薬剤であると判断され，「本剤以外の P2Y12 受容体拮抗薬などの抗血小板剤の投与が副作用の発現などにより困難な場合」との制限が設けられた上で承認された。

複合評価項目を用いることで多重性の問題を回避することが出来るが，主要評価項目に設定した複合評価項目の妥当性を検討する必要がある場合は，その構成要素や，その他に設定された複合評価項目，事後的に設定した複合評価項目など，様々な視点からの解析は重要である。

【参考文献】

[1] チカグレロル 審議結果報告書・審査報告書（2015 年 07 月 03 日）. Available online at: https://www.pmda.go.jp/drugs/2016/P20161026001/670227000_22800AMX00680_A100_1.pdf.pdf
[2] チカグレロル CTD. Available online at https://www.pmda.go.jp/drugs/2016/P20161026001/index.html

4.4. サブグループ解析における多重性の問題

4.4.1. Imlygic

IMLYGIC は，悪性黒色腫を対象疾患として，2015 年 12 月に欧州で初めて承認された腫瘍溶解性ウイルス／遺伝子治療用製品である。悪性黒色腫細胞にのみ感染する単純ヘルペス 1 型ウイルス遺伝子に，GM-CSF（顆粒球マクロファージコロニー刺激因子）遺伝子が組み込まれたものである。FDA では 2015 年 10 月に承認された。

本事例は，検証試験で全体集団における主要評価項目の優越性が検証されたが，事後のサブグループ解析では，病期が早期の被験者において治療効果が顕著であり，病期が後期の被験者では治療効果が見られなかった。この結果について審査過程で追加解析および議論が行われ，最終的に indication（いわゆる添付文書上における効能・効果）は，病期が早期の被験者だけに限定されることとなった。そして，添付文書（SmPC）に多重性の調整が行われていない事後の探索的なサブグループ解析が掲載された稀有な事例として紹介する。

申請時の Indication 案：

Talimogene laherparepvec is indicated for the treatment of adults with melanoma that is regionally or distantly metastatic.

承認時の Indication：

Imlygic is indicated for the treatment of adults with unresectable melanoma that is regionally or distantly metastatic (Stage IIIB, IIIC and IVM1a) with no bone, brain, lung or other visceral disease.

検証試験（005/05）において，主要評価項目である持続的奏効率（DRR）において優越性が検証されたが，主要な副次評価項目として最初に検定された全生存期間（OS）については，優越性が検証されなかった。

表 4.5.1-1　検証試験の結果（主要評価項目と主要副次評価項目）

	Study endpoint	Imlygic N = 295	GM-CSF N = 141
Durable response rate	Primary	16.3% (n = 48) (95% CI: 12.1, 20.5)	2.1% (n = 3) (95% CI: 0.0, 4.5)
		Odds ratio 8.9; (95% CI: 2.7, 29.2) P < 0.0001	
Overall response rate (% CR, % PR)	Secondary	26.4% (n = 78) (95% CI: 21.4%, 31.5%) (10.8% CR, 15.6% PR)	5.7% (n = 8) (95% CI: 1.9%, 9.5%) (0.7% CR, 5% PR)
Overall survival	Secondary	Median 23.3 (95% CI: 19.5, 29.6) months	Median 18.9 (95% CI: 16.0, 23.7) months
		HR: 0.79; (95% CI: 0.62, 1.00) p = 0.051	
Duration of response (ongoing response at last tumour evaluation)	Secondary	Not reached (Range: > 0.0 to > 16.8 months)	Median 2.8 months (Range: 1.2 to > 14.9 months)
		HR: 0.46; (95% CI: 0.35, 0.60)	
Time to response (median)	Secondary	4.1 months	3.7 months
Time to treatment failure (median)	Secondary	8.2 months (95% CI: 6.5, 9.9)	2.9 months (95% CI: 2.8, 4.0)
		HR: 0.42; (95% CI: 0.32, 0.54)	

[Imlygic SmPC Table 4 より転載]

事後的なサブグループ解析の結果，主要評価項目である持続的奏効率（DRR），副次評価項目である全奏効率（ORR），全生存期間（OS）において，病期が早期（Stage IIIB/IIIC/IV M1a）の被験者では，対照群より良好な結果が観察されたものの，病期が後期（Stage IV M1b/Stage IV M1c）の被験者では対照群とほとんど差がないことが示された。

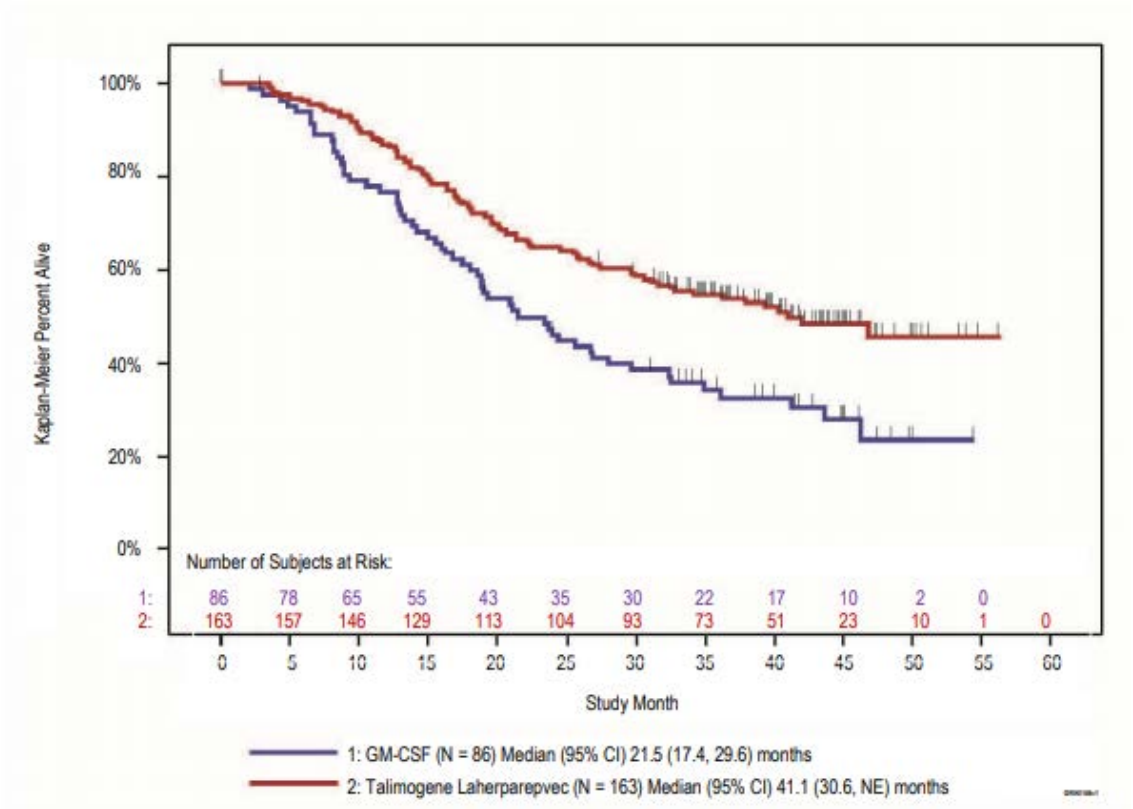
Table 5. Summary of results from exploratory subgroup analysis from Imlygic study 005/05

	DRR, (%)		ORR, (%)		OS (hazard ratio)
	Imlygic	GM-CSF	Imlygic	GM-CSF	Imlygic vs GM-CSF
Stage§ IIIB/IIIC/ stage IVM1a (Imlygic, n = 163; GM-CSF, n = 86)	25.2	1.2	40.5	2.3	0.57, (95% CI: 0.40, 0.80);
Stage§ IVM1B/ IVM1C (Imlygic, n = 131; GM-CSF, n = 55)	5.3	3.6	9.2	10.9	1.07, (95% CI: 0.75, 1.52);

§ American Joint Committee on Cancer (AJCC) staging 6th edition.

[Imlygic SmPC Table 5 より転載]

Figure 5. Kaplan-Meier estimate of overall survival by randomised treatment arm for disease stage IIIB/IIIC/ stage IVM1a (exploratory subgroup analysis)



[Imlygic SmPC Figure 5 より転載]

以上の結果をうけて，The Committee for Advanced Therapies（CAT）の要請により，Scientific Advisory Group meeting（SAG）が開催され，サブグループ解析の結果をどのように解釈するかについて議論がなされた。そして，SAG が，添付文書（SmPC）に全体集団とサブグループの両方の結果を（多重性を調整していないため，サブグループ解析は exploratory であることを明示したうえで）掲載することは有用であろうと回答しており，添付文書（SmPC）に全体集団とサブグループの両方の結果が反映されることとなった。

【参考文献】

[1] Imlygic Assessment Report. Available online at: https://www.ema.europa.eu/en/documents/assessment-report/imlygic-epar-public-assessment-report_en.pdf

[2] Imlygic SmPC. Available online at: https://www.ema.europa.eu/en/documents/product-information/imlygic-epar-product-information_en.pdf

5. 多重性調整に関する今後の課題

本章では試験計画の立案および試験結果の解釈において，非統計解析担当者と統計解析担当者の協働が必要となる事項を取り上げる。

5.1. 適切な仮説構造の選択

従来，検証試験では 1 つの主要評価項目を設定し，その主要評価項目に関して第 1 種の過誤確率を 5%以内に保った上で検定を行うことが一般的であった。近年，特に欧米においては，主要副次評価項目の結果を添付文書の効能・効果に記載することや，他剤との差別化を意図して，多重性を調整したもとでの検証された仮説を増やしたいという目的から，主要評価項目に加えて，主要副次評価項目について多重性を調整することがある。本邦では主要副次評価項目の多重性を調整している試験は多くないと考えられるが，国際共同治験の増加に伴って，本邦でも主要副次評価項目の多重性の調整を行う試験が増加する可能性がある。

なお，多重性を調整したもとで有意となった仮説を薬剤の付加価値とみなしうるかについては議論が必要である。例えば，主要副次評価項目について多重性を調整したもとで，予め検証目的で症例数設計された場合の検定結果と，そうでない場合の検定結果とが同一であったとしても，後者について「検証された」と見なせるかについては様々な見方があるであろう。

また，バイオマーカーを利用した開発も行われるようになってきており，バイオマーカーによって特に薬効が期待される部分集団と全体集団で多重性を調整することも行われるようになってきている。更に，開発の迅速化および効率化のために，検証試験において被験薬の複数の用量群の検討や中間解析の実施など，複数の多重性の要因が含まれる複雑な試験が計画されるようになってきている。

そのような試験目的（検証したい仮説）が複数存在する臨床試験での多重性の調整には，ゲートキーピング法などの複雑な仮説構造に対応した多重性調整方法が用いられることがある。これらの方法では，仮説の検定手順によっては検定手順の途中で有意ではない結果が出た場合に，それ以降の仮説が検討出来ない状況となる可能性がある。つまり，試験目的に含める仮説や，評価項目の臨床的な重要度といった仮説間の優先順位の設定により，得られる結果が異なる可能性がある。試験の目的を達成し，臨床的に解釈可能な結果を得るようにするためには，適切な仮説構造の選択が必要となる。仮説構造に含める検証したい仮説の特定，仮説間の優先順位を検討する際には，統計学的な観点だけでなく，疾患および薬剤の特性，疾患評価ガイドライン，医療現場のニーズ，規制当局の承認要件，開発戦略といった様々な観点からの検討が必要であり，開発計画の早期の段階から臨床開発部門のみならず薬事，マーケティング部門などの関係機能が協働して計画を進めていく必要性がある。

また，複数の地域での承認申請を考えている場合は，疾患評価ガイドライン，規制当局の承認要件の違いや添付文書の効能・効果の記載形式の違いにより，地域によって仮説構造が異なる可能性があることから，仮説構造については各規制当局とも協議し，合意しておくことが推奨される。非劣性が検証されたのち，事前に規定されていた検定順序にしたがって優越性を検証することは，検定に伴う多重性の問題は生じないことが知られている。ただし，非劣性仮説のみならず，優越性仮説についても検出力を考慮し，事前に必要被験者数を設定してある場合においても，多重性を考慮した検定手順により優越性まで有意であったとしても，「優越性仮説は検証された」とみなされない場合が存在する。したがって，仮説構造

のみならず想定される結果の解釈についても予め当局と議論しておくことが望ましい。

5.2. 症例数設計および多重性の調整方法の検討

1 つの主要評価項目を被験薬群と対照群の 2 群間で比較するといったような，検証したい仮説が 1 つしか存在しない検証試験では，その仮説が有意となる確率（検出力）を 80%～90%となるように症例数設計を行うことが一般的である。検証したい仮説が複数存在し，多重性の調整を行う場合には，症例数設計においても仮説構造および多重性の調整方法を考慮することがある。

下記に挙げるように，評価項目と用量の組み合わせで様々な試験成功の基準が定義可能であり，試験の成功確率の考え方（ある種の検出力のようなもの）も変わってくる。

- ✓ 「主要評価項目のみ有意になれば良い」
- ✓ 「主要評価項目に加えて主要副次評価項目も有意にする必要がある」
- ✓ 「被験薬の複数用量群と対照群の比較では，いずれか 1 つの用量群で有意になれば良い」
- ✓ 「すべての用量群で有意になれば良い」
- ✓ 「特定の用量（例えば，最高用量）で有意になりさえすれば良い」など

つまり，個々の検定の検出力にのみ注目するのではなく，試験目的を達成するために必要な仮説が有意となる確率を確保出来るように症例数設計を行う必要性について検討することは有益である。

多重性の調整を行うと多重性の調整を行わない場合と比較して，個々の検定は有意になりにくくなることがあるため，個々の検定で多重性の調整を行わない場合と同じ検出力を確保するために必要な症例数は増加することがある。必要症例数の過度な増加を抑えるためには，仮説構造を入念に定義して，試験目的達成に必要な仮説の検定に割り当てられる有意水準に関する重みを大きくすることで有意水準をある程度大きく確保することや，より検出力が高い多重性調整方法を選択することが考えられる。

なお，固定順序法で全ての仮説を有意水準 5%で検定する場合であっても，ある仮説が有意になるためには，それより前の仮説が全て有意でなければならないため，後の仮説になるほど多重性の調整を行わない場合と比べて有意になる確率は低下する。例えば，1 番目の仮説と 2 番目の仮説を個別に検定した場合の検出力がそれぞれ 90%であったとしても，固定順序法で 2 番目の仮説が有意になるには「1 番目の検定が有意かつ，2 番目の検定も有意」となる必要がある。2 つの検定の間に関連がなく独立である場合，2 番目の検定の検出力は約 80%（90%×90%=81%）となってしまう。固定順序法だけでなくゲートキーピング法などの階層的に検定を行う方法は，一般的に仮説構造の下層に位置する仮説ほど検出力は低下する。特に，仮説構造の途中に検出力が低い（つまり，あまり効果が見込めない）仮説がある場合，それ以降の仮説に進むことができる確率は低くなる。したがって，仮説間の優先順位の検討は，各仮説の臨床的な重要度のみならず検出力も考慮すると共に，症例数設計では試験の目的を達成するために必要な仮説のみを含めることが適切であると考えられる。

以上のように，多重性の調整方法および症例数設計に含める仮説の選択によっては，目標症例数が増加し，試験成功の確率が低下する危険や，試験の実施費用および開発期間の増大を招く可能性がある。また，必要以上の患者を試験に組み入れることの倫理的な問題もある。したがって，症例数設計に含める仮説と適切な多重性の調整方法の検討も慎重に行う必要があり，試験計画時に臨床開発部門，薬事およびマーケティングなどの関係機能の非統計解析担当者と統計解析担当者が協力しながら検討していく必要がある。

5.3. 多重性を考慮した意思決定の重要性

複数の多重性の要因が存在する場合は，多重性の調整方法も複雑となり試験結果の解釈を難しくすることがある。従って試験の計画の段階から仮説構造（検定手順と意思決定の方法）をきちんと理解しておく必要がある。意思決定の方法はその試験結果の解釈やその後の開発戦略に影響を与えるため，統計解析担当者のみならず，多くの機能に関わる問題となる。従って，臨床開発部門をはじめとして，薬事，マーケティング部門など関係機能と協働して検討する必要がある。

また，医療関係者へ情報提供する際も，試験成績が正しく理解されるように，どの様な多重性調整方法が用いられたかについて適切に提示する必要がある。2019 年に公開された医療用医薬品製品情報概要などに関する作成要領（解説付き）では，多重性調整に関しても製品情報概要の「解析計画」に記載するように示されている。

「解析計画」
医療関係者が試験成績を正しく評価できるよう、以下のような内容がプロトコールに事前規定されている場合は、試験デザインの解析計画に記載してください。中間解析やアダプティブデザインを採用した場合は解析計画について明示してください。

記載例： 解析の目的・回数・タイミング・有意水準 等
① 非劣性試験の検証的な解析結果から、対照薬に対する優越性を主張する場合（非劣性が検証された場合に、優越性の検証を行うこと）
② 主要評価項目に対する主要な解析と他の評価項目に対する解析を組み合わせ、これらの解析全体として検証と位置付けられている場合
上記①②のような枠組みにおいて、複数の検定を行う順序を定めることにより、検定の多重性問題に対処している場合（Gate Keeping Strategy、閉検定手順等）は、検定の順序も明示してください。
③ 年齢・性別等、背景因子による層別解析を行う場合
④ 試験に組み入れられた被験者の部分集団に対する解析（サブグループ解析；日本人集団の解析、重度の肝機能障害を有する集団の解析等）を行う場合

探索試験では検証試験とは異なり，多重性の厳密な調整は規制上要求されていない。しかし，試験結果の解釈において多重性の影響を考慮せずに意思決定を行うと，探索試験の結果を過大評価しその後の開発を進めてしまうことで，その後の試験が失敗する原因となる可能性がある点には注意が必要である。検証試験における探索的な解析結果も同様であり，探索的な解析結果を有効性の検証的な根拠として主張することは出来ない。仮説生成を目的とした探索的解析結果の解釈においても多重性の観点については注意が必要である。

多重性は臨床試験における様々な場面で生じる可能性がある。結果の解釈において多重性をどのように考慮するかは，解析結果がどのような目的で用いられるかによって変わってくる。解析結果を適切に解釈するためには，試験目的の共通理解の下，臨床的な観点と統計的な観点の両方が必要であり，非統計解析担当者と統計解析担当者が適切にコミュニケーションを取り，多重性を考慮した意思決定の枠組みを入念に議論し共通認識を持っておくことが重要である。

【参考文献】

[1] ICH. (1998). ICH harmonised tripartite guideline: statistical principles for clinical trials-E9. Available online at: https://database.ich.org/sites/default/files/E9_Guideline.pdf
[2] CPMP. (2002). Points to consider on multiplicity issues in clinical trials. Available online at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003640.pdf
[3] FDA. (2017). Multiple endpoints in clinical trials: draft guidance for industry. Available online at: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM536750.pdf
[4] 日本製薬工業協会 医療用医薬品製品情報概要審査会. (2019). 医療用医薬品製品情報概要等に関する作成要領（解説付き）. Available online at: http://www.jpma.or.jp/about/basis/drug_info/pdf/drug_info05.pdf

6. おわりに

本報告書では，検定（あるいは，信頼区間を検定のように用いた検討）に伴う多重性の問題，ICH ならびに各規制当局の考え方，多重性を考慮した臨床試験の実例および審査において問題となった事例などについて触れた。これらを踏まえ，多重性の問題は試験計画時から試験結果の解釈およびその後の開発計画に至るまで，多くのステークホルダーを巻き込んだ広範にわたる問題であることを示した。多重性の問題は統計解析担当者のみで対処すべき問題ではなく，医薬品開発に関わる多くの関係機能に関わる問題であることは重ねて強調したい。

本報告書をきっかけに，多重性の問題に関する興味・理解が深まり，臨床試験に関わる多くの担当者が試験計画時から統計解析担当者と議論し，必要に応じて規制当局とも適切にコミュニケーションを取ることで，開発計画および試験目的に即した理想的な試験が実施・解釈されることを期待する。

資料作成担当者

日本製薬工業協会 医薬品評価委員会 データサイエンス部会 タスクフォース 4

多重性問題検討サブチーム

- 伊庭　克拓　　大塚製薬株式会社
- 浦田　正夫　　サノフィ株式会社（2017 年 4 月まで）
- 小川　直之　　株式会社三和化学研究所（2017 年 4 月より）
- 神浦　俊文　　日本新薬株式会社
- 菅波　秀規　　興和株式会社*
- 土屋　悟　　大日本住友製薬株式会社*
- 冨金原　悟　　小野薬品工業株式会社*
- 森田　祐介　　杏林製薬株式会社（2019 年 7 月まで）
- 吉田　征太郎　中外製薬株式会社

*タスクフォースリーダー兼，推進委員

以上

Appendix 1　各規制当局間の共通点および相違点のまとめ

ICH E9 (1998)	CPMP (2002)	CHMP (2017)	FDA (2017)
＜主要評価項目＞			
✓ 主要評価項目の個数は出来るだけ１つにする。 ✓ 複数の主要評価項目を用いる場合には第１種の過誤を制御する方法を治験実施計画書および統計解析計画書に記載。 ✓ 複数の主要評価項目を用いる場合に第２種の過誤が増大し，それによって必要な被験者数が増加する。			
✓	✓ 複数の主要評価項目が必要な場合があり，closed testing procedure に属する方法を用いれば有意水準の調整が不要であること，closed testing procedure での検定と整合する信頼区間が構成できる。		✓ 複数の主要評価項目が必要な場合，開発の早期段階から十分に評価することで，後の検証的な試験で単一の主要評価項目を用いることの助けになる。
＜多重性の調整＞			
✓ 多重性を回避あるいは減じること。	✓ 治験実施計画書や解析計画書に多重性の調整方法を事前記載すること。		
✓ 多重性が存在する場合，調整は常に考慮すべきであり，調整方法の詳細，又はなぜ調整は必要ないと考えるのかという説明を統計解析計画書に記載。	✓ 多重性を調整する方法を選択する際に検定結果と整合する信頼区間を構成できるかどうか，結果について臨床的な解釈ができるかどうかの考慮が必要。		
＜安全性評価項目＞			
✓ 安全性および忍容性の変数に対して検定を用いる場合，通常は第１種の過誤よりも第 2 種の過誤により注意を払うべき。	✓ 安全性の変数であってもその検定結果が承認や labeling claim に反映される場合には主要変数と同様に多重性の問題を扱うべき。		
＜副次的評価項目＞			
✓ 副次評価項目について，その数を試験で答えるべき限られた少数の問題と関連して制限すべき。	✓ 副次評価項目が添付文書などへの効能効果への記載を意図しない場合にはその信頼区間や検定結果は探索的に用いられる。		
	✓ 主要評価項目で有効性が示された場合にのみ副次評価項目の検証が可能。		
＜多重性調整方法＞			
	✓ 一般的でない方法を使う場合には規制当局と相談とすること。		
	✓ 臨床推奨用量を決めるための用量反応試験に対しては FWER を制御。	✓ 用量反応試験では検定よりも推定に基づくべき。 ✓ 有意な最低の用量を見出すという手順は結果の解釈を誤ることがある。	
指定した全ての主要変数において有効性を示すことが試験の目的（co-primary endpoint）である場合，第１種の過誤を調整する必要はない。 しかし第２種の過誤および必要な被験者数への影響は慎重に考慮すべき。			
＜composite endpoint＞			
	✓ composite endpoint を rating scale と生存時間解析に関連するものの２種類に区別。		✓ 複数の要素から成り立っている評価項目を multi-component endpoint と

ICH E9 (1998)	CPMP (2002)	CHMP (2017)	FDA (2017)
			呼び，その中でもイベント発現までの時間を解析対象とする場合のみを composite endpoint とする。
			✓ composite endpoint の各要素は個別に解析して常に試験の報告書に含めるべき。
＜推定における多重性＞			
		✓ 検定に伴う多重性の問題のみではなく，推定における多重性の問題について言及。	