

臨床試験計画書を作成する上での留意点一覧

2020年3月31日

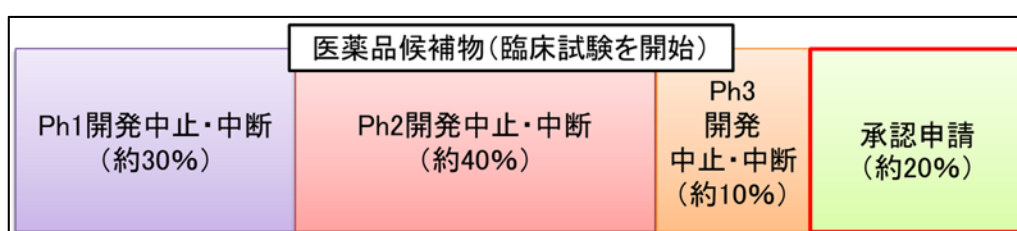
AMED 医薬品等規制調和・評価研究事業

ネガティブ試験等の結果を用いた効率的な試験デザイン等の検討に関する研究

<本研究の背景及び目的>

医薬品開発における成功確率は、非臨床から臨床の段階に移行した医薬品候補化合物においても、決して高くないことが報告されている。

また、臨床試験の各段階における開発中止・中断の割合は、下図の通りである。



これらのうち、第Ⅰ相臨床試験、第Ⅱ相臨床試験の段階で中止・中断となった医薬品候補化合物は、医薬品としてのポテンシャルが足りなかった可能性があり、試験デザイン等を工夫しても、有効性等を示せない可能性がある。一方、第Ⅲ相（ピボタル）臨床試験の段階に至った医薬品候補物については、第Ⅱ相臨床試験までに有効性が示唆され、ある程度の安全性の確認がされていることから、この段階において中止・中断となった医薬品候補化合物については、試験デザイン等の問題により、ピボタル試験で有効性及び安全性を示すことができなかったものが一定数存在する可能性がある。

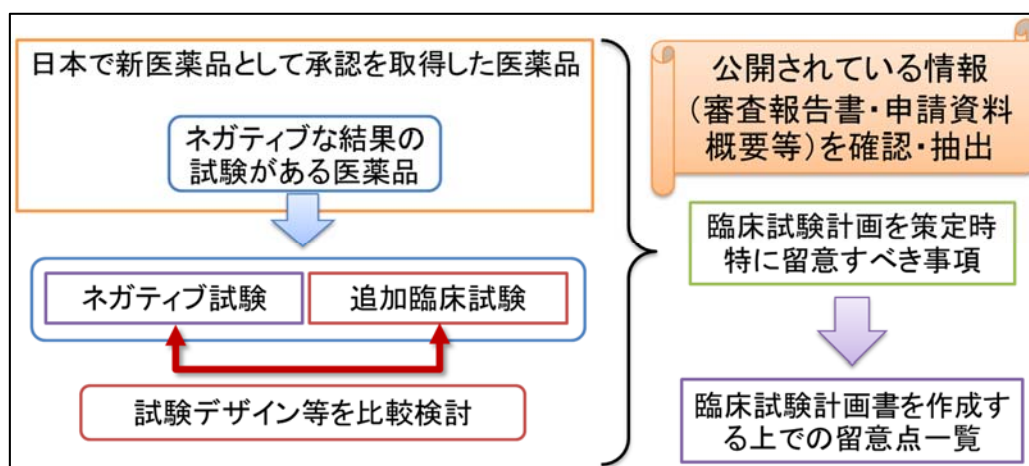
そこで我々は、これらの医薬品候補化合物のうち、日本で新医薬品として承認を取得した医薬品であり、臨床データパッケージにネガティブな結果の試験がある医薬品に焦点を当て、情報が入手可能である承認審査資料を基に、ネガティブな結果となってしまった要因等を検討した。

上記の検討結果を基に、医薬品の臨床開発時、特に重要な臨床試験（いわゆる、ピボタル試験）の計画時に留意すべき事項を明らかにすることを目的とした。

<本研究の内容>

本研究では、2005年～2017年に日本で新医薬品として承認された医薬品1,324品目を対象に、当該医薬品の承認申請の臨床データパッケージの中で、有効性及び安全性の評価を行うために、重要な臨床試験（いわゆる、ピボタル試験）のうち、対照群を設定した比較試験で、ネガティブな結果を有する臨床試験（ネガティブ試験）の情報を収集し、検討した。なお、1990年代に開始された試験は除外した。

以下に、本研究で検討対象とした臨床試験、ネガティブ試験の定義を示す。



<検討対象の臨床試験>

- ・ 国内又は日本を含む国際共同のピボタル試験であり、対照群を設定した比較試験
- ・ (有効性を示す主な根拠となる試験)：新医薬品の承認申請の臨床データパッケージの中で、有効性及び安全性の評価を行うために、重要な試験
 - 1) 通常、第Ⅲ相臨床試験 (第Ⅱ/Ⅲ相臨床試験も含む)
 - 2) 抗悪性腫瘍剤、希少疾患等において、第Ⅲ相臨床試験が実施されていない場合、第Ⅱ相臨床試験
 - 3) ブリッジング戦略を用いている場合、ブリッジング試験 (通常、第Ⅱ相臨床試験)

<ネガティブ試験の定義>

以下の 1) 又は 2) を満たす。

- 1) 主要評価項目の主要解析において、有意差、非劣性、用量反応性等が示されていない試験
- 2) 医薬品医療機器総合機構 (PMDA) の審査・相談等により、承認する上で、有効性、安全性等が不十分であると判断された試験。

新医薬品の承認品目数	ネガティブ試験を有する新医薬品目数	ネガティブ試験数
1,324 品目	62 品目	59 試験 [#]

ネガティブ試験数は、70 試験あったが、1990 年代に開始された試験 11 試験は除外

当該研究結果を踏まえ、通常、医薬品の臨床試験実施計画書（プロトコル）に含まれる以下の①～⑨の項目毎に、本研究班において、検討したネガティブな結果を有する臨床試験成績を踏まえた臨床試験計画書を作成する上での留意点を一覧に示した。

- ① 対象患者
- ② 適格基準
- ③ 目標症例数
- ④ 治験デザイン
- ⑤ 併用薬・併用療法
- ⑥ 治験薬、用量及び投与方法
- ⑦ 評価項目
- ⑧ 統計手法
- ⑨ その他（治験の実施時期等）

留意点の一覧は、「項目」、「留意点」、「具体的な事例」で構成した。また、「具体的な事例」には、「疾患領域」、「適応」、「ネガティブな結果となった原因」、「関連項目」で構成した。

「具体的な事例」で取り上げた品目において、臨床試験がネガティブな結果となった原因が複数存在する品目もあることから、同じ品目において、他の項目にも記載がある場合、「関連項目」に他の項目の記載場所を明示した。

以下の一覧においては、ICH E8「臨床試験の一般指針」、E9「臨床試験のための統計的原則」、E10「臨床試験における対照群の選択」の記載も参考として提示した上で、留意事項を示した。

研究開発代表者

岐阜大学医学部附属病院 浅田 隆太

研究開発分担者

東京大学大学院薬学研究科 小野 俊介

東北大学大学院医学系研究科 山口 拓洋

国立精神・神経医療研究センター 中村 治雅

名古屋大学医学部附属病院 清水 忍

① 対象患者

	項目	留意点	具体的な事例			
			疾患領域	適応	ネガティブな結果となった原因	関連項目
①-1	評価に適していない対象患者の設定	対象患者について、いくつかのタイプがある場合 (例：通年性/季節性、急性/亜急性 等)、医薬品の作用機序、評価項目等を踏まえて、十分な均質性を維持するために、一定の範囲内で評価が可能な対象患者を選定できているか、十分に検討する必要がある。	アレルギー	アレルギー性鼻炎	通年性アレルギー性鼻炎患者では、日常的に鼻症状が現れていることから、鼻症状が比較的軽症の患者では症状の変化を精度良く捉えられなかった可能性がある。(追加臨床試験では、季節性アレルギー性鼻炎患者を対象として実施)	なし

② 適格基準

	項目	留意点	具体的な事例			
			疾患領域	適応	ネガティブな結果となった原因	関連項目
②-1	自覚症状に関連する主要評価項目における評価に適切な集団の選定が不十分	自覚症状に関連する評価項目を主要評価項目とする場合、選択・除外基準にベースラインの症状の重症度、評価項目に関与する可能性がある合併症・症状等、観察期に主要評価項目を複数回測定し、そのバラツキに関する規定（バラツキが小さい等）を設定する必要があるか、十分に検討する必要がある。	アレルギー	アレルギー性鼻炎（小児）	一般的に小児では成人と比べて自覚症状の把握が難しいことから、特に症状が軽度の被験者では、症状の変化を精度よく捉えられなかった可能性がある。	⑦-3
			アレルギー	アレルギー性鼻炎（小児）	通年性アレルギー性鼻炎合併症例は、常に抗原に曝露されていることから、非合併症例よりも鼻症状スコアのベースラインの値が高い傾向にあり、花粉曝露による鼻症状の一時的な変化を感度よくとらえることができない可能性がある。	⑦-4
			精神・神経用	不眠症	主要評価項目である自覚的睡眠潜時*について、評価がよりの確に実施できていると考えられる集団を選定できていなかった。 *睡眠潜時：就床から入眠までの時間	なし
			精神・神経用	社会不安障害	投与開始時 LSAS*合計点が 30 点未満であった患者は、海外試験と比較して多く、軽症の患者が多く組み入れられたため、計画時に想定していた LSAS 合計点の変化量（本剤群で最大 30 点）が得られず、プラセボ群に対する優越性が示せなかった。 *LSAS（Liebowitz Social Anxiety Scale）：社会不安障害の臨床症状を図るための尺度。「行為状況」に関する 13 項目と「社交状況」に関する 11 項目の全 24 項目からなる。	⑥-1

②-2	薬剤の投与対象として、適切な集団の選定が不十分	開発薬の作用機序等を考慮して、有効性を評価する上で、適切な集団を選定するために、選択・除外基準に病態の進行度・重症度、罹患期間、症状の慢性化の程度、評価項目に関与する可能性がある合併症・症状等を考慮する必要があるか検討する必要がある。	精神・神経用	筋萎縮性側索硬化症	追加解析において、確実な診断をされ、軽症（登録時での病勢進展が軽微で呼吸機能が正常）、かつ発症後2年以内の患者において、有効性を確認できると考察されたことから、評価に適切な集団が選定されていなかった。	なし
			精神・神経用	統合失調症における精神運動興奮における機能障害	投与対象として必ずしも適切ではないと考えられる「非精神病性の興奮と考えられる Acute Agitation を有する患者」がプラセボ群に、「治療抵抗性で慢性的に高い PANSS-EC*合計点を示す慢性期統合失調症の固定的な症状を有する患者」が実薬群にそれぞれ偏在していた。 *PANSS-EC (Positive and Negative Syndrome Scale-Excited Component) : 統合失調症の精神症状の評価尺度	④-1
②-3	適格基準を満たす被験者数の想定が不十分	既承認の薬剤の新用法・用量（用法・用量：小児の用法・用量の追加等）に関する臨床試験の場合、臨床試験の対象患者は当該薬剤の投与経験がある可能性があることから、選択・除外基準の設定の際、当該薬剤に関する規定を十分に検討する必要がある。（例：「当該薬剤の服用経験がない」ではなく、「当該薬剤を○週以内に服用していない」等）	精神・神経用	強迫性障害（小児）	強迫性障害について、成人用量は承認されており、小児患者に最も多く使用されている状況であったことから、選択基準である「本剤の服用経験がない症例」に該当する小児被験者は少なく、被験者登録は困難であった。	なし

ICH E9「臨床試験のための統計的原則」

2.2.1 対象集団

検証的試験では一般に、試験治療の効果を精度よく推定するために十分な均質性を維持する一方、目標集団の範囲内でできるだけ広い患者を対象とするような選択基準及び除外基準を定めることが有益である。

ガイドラインには、上記の通り、試験治療の効果を精度よく推定するために十分な均質性を維持する選択基準及び除外基準を定めることが有益と記載されているものの、具体的な事例等の記載がなされていないことから、上記の一覧に示した留意点が浮かび上がってきたと考える。

ガイドラインの記載とともに、具体的な事例を示すことにより、ガイドラインが有効に活用可能となると考える。

③ 目標症例数

	項目	留意点	具体的な事例			
			薬効分類	適応	ネガティブな結果となった原因	関連項目
③-1	臨床試験結果の群間差と目標症例数の設定に用いた群間差の差の違	臨床試験結果における群間差が目標症例数の設定に用いた群間差より小さかったため、有効性を示すことができなかった事例が散見されることから、対照群（プラセボ群）の効果も含め、保守的に見積もることが必要な場合がある。	解熱・鎮痛・消炎	高安動脈炎	計画時に無再発率の実薬群とプラセボ群の群間差を過大に設定したことが原因と考えられた。	なし
			循環器系	頻脈性心房細動	計画時に想定したよりも用量群間の効果の差が小さく、ばらつきが大きかった。	なし
			皮膚疾患用	尋常性ざ瘡	対照群の「総皮疹数のベースラインからの減少率」が81.6%であり、目標症例数の設定根拠に用いた国内第Ⅱ相試験（58.3%）、海外第Ⅲ相試験（35.6～48.2%）と比較して想定外に高かった。	なし
			消化管用	潰瘍性大腸炎	海外試験等を参考に、実薬群のUC-DAI*スコアは対照薬群よりも0.5上回って改善すると想定していたが、実薬群と対照薬群のUC-DAI*スコアの変化量の群間差の調整済み平均値 [95%信頼区間] は、0.3 [-0.5, 1.1] と想定より小さかった。 *UC-DAI (Ulcerative Colitis Disease Activity Index) : 潰瘍性大腸炎の疾患活動指数	なし
			眼科用	緑内障、高眼圧症	既存治療に対する実薬の併用効果が想定よりも小さかったことによる検出力不足である可能性が考えられた。	なし
			精神・神経用	うつ病・うつ状態	プラセボ群のHAM-D*17合計スコアの変化量が想定より大きかった。 *HAM-D (Hamilton Depression Rating Scale) : ハミルトンうつ病評価尺度	なし
			精神・神経用	てんかん患者の部分発作	プラセボ群における部分発作減少率が前相である第Ⅱ/Ⅲ相臨床試験より高かった。	⑥-1

③-2	目標症例数の設定に用いた対象集団と計画する臨床試験の対象集団との違い	目標症例数の設定に用いた対象集団と計画する臨床試験の対象集団について、有効性に影響を与える可能性のある因子が試験間で割合が異なるため、期待していた有効性を示すことができなかつた事例があることから、計画している臨床試験の対象集団が目標症例数の設定に用いた臨床試験の対象集団と、重症度別の割合、併用薬の種類、有無の割合等が異なるような方を検討する必要がある。(割付因子等)	呼吸器系	慢性閉塞性肺疾患	海外試験の群間差（努力呼気 1 秒量トラフ値の変化量：67 及び 69mL）から、本試験の群間差を 80mL と推定した。しかし、海外試験と比べて、可逆性を有する患者の割合が高かつたことから、海外試験よりも効果が高く見込まれる集団が多く組み入れられた。	なし
			精神・神経用薬	てんかん患者の部分発作	本試験では、症例数を設計に用いた海外試験と比較して、症候性てんかん、抗てんかん薬の併用数が 2 剤以上の患者の割合が高く、治療反応性が不良の患者が多く組み入れられた可能性があつた。	なし
			血液製剤	水疱性類天疱瘡	本試験では重症の患者が登録されたことから、目標症例数の算出に使用した「PDAI*を用いたスコア」の各群共通の標準偏差が、第Ⅱ相試験（16.6）と比べて、27.1 と大きくなつた。 *PDAI（Pemphigus Disease Activity Index）：天疱瘡重症度判定基準	なし
			精神・神経用薬	アルツハイマー型認知症	介護サービスの利用割合（36.8%）は、前相の国内試験（25.1%）及び海外試験（10.2%及び 25.2%）より、本試験で高かつた。	なし
			精神・神経用薬	アルツハイマー型認知症	2006 年 4 月の介護保険制度の改正に伴い、介護サービスの利用が増えた。CIBIC plus-J*では、介護サービス利用なしの集団で実薬群とプラセボ群との間に有意差が認められたが、利用ありの集団では有意差が認められなかつた。 *CIBIC plus-J（Clinician’s interview-based impression of change plus Japanese version）：全般的臨床症状評価日本版	なし
			精神・神経用薬	アルツハイマー型認知症	デイケア・デイサービスを利用している場合、CIBIC plus-J の評価者は、実際にみられている臨床症状の変化を十分に捉えることができなかつた。本試験ではデイケア・デイサービスの利用率（55.3%）が高く、CIBIC plus-J では有意な差を検出することができなかつた。	なし

③-3	実施可能性の検討が不十分	目標症例数を設定する際、実施可能性を考慮した上で、設定した登録期間内に組入れ可能な例数であるか検討する必要がある。	精神・神経用薬	斜視	10PD*以上 20PD 未満の斜視患者の集団では、①斜視による日常生活への影響が少ないため治験に参加してまで治療しようと思わないこと、②就学・就労により多忙であるため頻回な通院が困難であること、③眼科愁訴がない患者は眼科を受診しない、20PD 未満の斜視は整容面で気にならない、小角度の斜視による複視の解消にはプリズム眼鏡の着用が有用である等の理由により治療対象患者が少ないことから、組入れが進まなかった。 *PD (プリズム ジオプトリー) : 遠見斜視角と近見斜視角の平均値	なし
-----	--------------	---	---------	----	---	----

ICH E9 「臨床試験のための統計的原則」

3.5 必要な被験者数

被験者数の計算は、主要な解析で用いる解析対象集団に基づくべきである。解析対象集団が「最大の解析対象集団」である場合、効果の大きさに関する見積値は、治験実施計画書に適合した対象集団（用語集参照）の場合に比べて小さくする必要があろう。これは、試験治療を中止した被験者又は服薬遵守状況の悪い被験者を解析に含めることにより、試験治療の効果が薄められることを考慮するためである。このときばらつきに関する仮定も再検討する必要がある。

ガイドラインには、上記の通り、「効果の大きさに関する見積値は、治験実施計画書に適合した対象集団（用語集参照）の場合に比べて小さくする必要があろう」と記載されているものの、症例数設計の際、どの程度、保守的に見積もるか等、判断が難しいことから、上記の一覧に示したような具体的な事例を示すことは意味があると考える。

④ 治験デザイン

	項目	留意点	具体的な事例			
			薬効分類	適応	ネガティブな結果となった原因	関連項目
④-1	主要評価項目に関連する項目に関する投与前値（前観察）の群間のバラツキ	主要評価項目に関連する因子が群間で異なることにより、結果に影響を及ぼす可能性があることから、選択・除外基準において規定、又は割付因子にする必要がないか十分に検討する必要がある。	アレルギー	アレルギー性鼻炎	観察期の総合症状スコア（主要評価項目）は、実薬で他群と比較して低く、治療による症状軽減の余地が少なく、スコア変化量が他群と比較して小さくなった可能性があった。	⑦-4
			解熱・鎮痛・消炎	関節リウマチにおける関節局所の鎮痛	投与群間における治験薬貼付開始時の VAS 値（主要評価項目）の影響があると考えられた。	⑤-1
			消化管用	非代償期の慢性膵炎又は膵切除	観察期の脂肪吸収率（主要評価項目）に群間に偏りがみられた（プラセボ群 46.2%、低用量群 58.1%、高用量群 61.5%）。	なし
			精神・神経用	統合失調症における精神運動興奮における機能障害	プラセボ群に「非精神病性の興奮と考えられる Acute Agitation を有する患者」、実薬群に「治療抵抗性で慢性的に高い PANSS-EC 合計点を示す慢性期統合失調症の固定的な症状を有する患者」がそれぞれ偏在した。	②-2
④-2	プラセボ効果を排除するための対応が不十分	治療介入による期待が主要評価項目の結果に影響を及ぼす可能性がある領域では、その影響を排除する必要がないか検討する必要がある。	精神・神経用	うつ病・うつ状態	いずれの投与群でも、特に投与開始後 1 週間の HAM-D17 合計評点の変化量がその後の期間に比べて大きかった。この原因として、治験開始時より治験コーディネーターが診療に関わることによる環境変化やプラセボ効果が考えられた。（追加試験では、プラセボリードイン期間を設定）	⑥-1

④-3	対照群が不適切	シュードプラセボを用いる場合は、用量が適切（有効性を示さない量）であるか、作用機序等を踏まえて、十分に検討する必要がある。	精神・神経用	うつ病・うつ状態	海外における最小有効用量の 1/4 量を対照群として設定した。しかし、海外試験のプラセボ群より、有効率が大きく、計画時の想定より大きかった。	なし
-----	---------	---	--------	----------	--	----

ICH E9 「臨床試験のための統計的原則」

2.3.2 ランダム化（無作為化）

基準となる時点で測定された重要な予後因子（例えば、疾患の重症度、年齢、性、など）で層別することは、層内でバランスのとれた割付を促進するために有益とってよい。

ガイドラインには、上記の通り、層別について、疾患の重症度、年齢、性を例に挙げているが、上記の一覧では、ベースライン時の主要評価項目に関する項目の群間における偏りが影響した例が多くみられている。ガイドラインに明確な記載がないことから、一覧で具体的な事例を示すことは、ガイドラインを活用の助けとなると考える。

⑤ 併用薬・併用療法

	項目	留意点	具体的な事例			
			薬効分類	適応	ネガティブな結果となった原因	関連項目
⑤-1	有効性に影響を与える可能性がある薬剤の併用	同様の作用を示す可能性のある薬剤・療法等を併用する場合、有効性評価に影響を与える可能性があることから、併用禁止薬・禁止療法又は併用制限薬・制限療法とすることを検討する必要がある。	解熱・鎮痛・消炎	関節リウマチにおける関節局所の鎮痛	全身性 NSAIDs 製剤を併用したことにより貼付部位への鎮痛評価へ影響した可能性がある。	④-1

⑥ 治験薬、用量及び投与方法

	項目	留意点	具体的な事例			
			薬効分類	適応	ネガティブな結果となった原因	関連項目
⑥-1	実薬群における有害事象による早期中止の影響	治験薬投与開始の早期において、有害事象等による中止が多かったことが有効性に影響を及ぼす可能性があることから、治験薬の用量及び投与方法について、特に、開始用量、漸増期間、漸増の用量幅等を十分に検討する必要がある。	精神・神経用	てんかん患者の部分発作	漸増期（投与開始後2週まで）に投与中止に至った有害事象（プラセボ群0例、実薬群6例）が実薬群で多く認められたため、実薬群の部分発作発現頻度変化率が想定よりも小さくなった可能性があった。	③-1
			精神・神経用	うつ病・うつ状態	実薬群として40mg群と60mg群（40mgから開始し60mgへ増量）を設定した。投与開始後1週未満の有害事象中止例が、実薬群で、シュードプラセボ群と比較して多かったことから、効果が十分に発揮される前に、投与中止に至った症例の割合が高くなり、適正な薬効評価が出来なかったことが原因の一つと考えられた。	④-2
			精神・神経用	社会不安障害	実薬20mg群：10mgから開始、40mg群：20mgから開始し、1週ごとに10mgずつ増量する計画であった。 40mg群では、有害事象発現による早期脱落が多く、早期脱落例（治験薬投与後3週間以内）を除外した場合にはプラセボ群に対する有意差が認められており、早期脱落例が有効性評価に影響を及ぼした可能性が考えられた。	②-1
			呼吸器系	特発性肺線維症	漸増期間を7日間と設定していたが、10/73例が有害事象（胃腸障害、投与初期に集中）のため、最高用量まで増量できなかった。	なし

⑥-2	治験薬の用量調整の制限による影響	治験薬の用量及び投与方法について、厳格な用量調整基準を設けることにより、有効性を示すことができないことがあることから、基準の必要性等について、十分に検討する必要がある。	持続型赤血球造血刺激因子製剤	腎性貧血	厳格な用量調整基準を設けて目標値内での医師の裁量による用量調整を制限した。その結果、投与開始から目標とする Hb 濃度に到達するまでの期間（中央値 [95%信頼区間]）が実薬群でより長く要したため（実薬群 49 日 [39 日, 56 日] 及び対照薬群 28 日 [21 日, 28 日]）、実薬群では目標 Hb 濃度付近における用量調整の機会が少なくなった。なお、主要評価項目は、評価期間における目標 Hb 濃度維持率であった。	なし
⑥-3	治験薬の用量が不十分	検証試験における検討用量が有効性を示すのに十分な用量であるか慎重に検討する必要がある。	ワクチン	新型インフルエンザ (H5N1)	<p>HA*含量 15μg では、非劣性を示すことができず、抗原量が不足していた可能があった。</p> <p>注) 新型インフルエンザ発生時の感染拡大防止のため、限られた量の抗原を用いて、より多くの国民にワクチンを迅速に提供することが重要であることから、本試験では、1 回接種当たりの抗原量を減らすことを目的に実施されており、特殊な状況であることに留意する必要がある。</p> <p>*HA (Hemagglutinin) : インフルエンザウイルス表面に存在する抗原の亜型</p>	なし

⑥-4	評価に影響を与える可能性がある治験薬（プラセボ）の投与回数が不十分	自覚症状に関連する評価項目を主要評価項目とする場合、患者個々のデータの精度を高めるために、投与回数を多く設定することが有用な場合があるので、検討する必要がある。	解熱・鎮痛・消炎	癌患者における突出痛	二重盲検期では6回の突出痛（本剤4回、プラセボ2回）に対し、突出痛毎に実薬又はプラセボをバックル投与とし、投与順序は各患者に無作為に割付けした。海外第Ⅲ相試験では実薬及びプラセボの投与回数は、それぞれ7及び3回（計10回）であったのに対し、本試験では実施可能性を考慮し、プラセボの投与回数を可能な限り少なくするため、それぞれ4及び2回（計6回）と設定しており、二重盲検期におけるプラセボの投与回数が少なかったことが患者評価に影響を与えた。	なし
⑥-5	治験薬の投与期間が不十分	治験薬の用法を決める上で、薬剤の作用機序、病態による適切な投与タイミング、回数等を検討することが必要である。	抗ウイルス	A型又はB型インフルエンザウイルス感染症の予防	インフルエンザウイルス感染症患者（初発患者）と同居する家族又は共同生活者を対象に、1週間に1回投与していたが、初発患者のウイルス排出量が多いと考えられる2及び3日目に投与を設定していなかった。	なし

⑦ 評価項目

	項目	留意点	具体的な事例			
			薬効分類	適応	ネガティブな結果となった原因	関連項目
⑦-1	評価方法・判定基準の統一が不十分	より客観的な評価を行うために、中央判定等の第三者判定を行う場合にも、評価方法、手順、基準を予め設定し、評価者間でバラツキが少なくなるように工夫することが必要である。	体内診断	MRI における造影	主要評価項目の評価は、3名の専門医が盲検下で読影していたが、評価方法が統一されておらず、評価にばらつきが生じた。	なし
			消化管用	逆流性食道炎	試験開始前に登録時の内視鏡判定で明確な基準を設定しなかったため、適切な判定が行われなかった可能性があった。（追加試験では、LA 分類*の画像を用いて撮影方法や写真の質の統一を図り、内視鏡判定委員会に提出する写真の治療期開始前での再確認を必須とした。） *LA（ロサンゼルス）分類：食道炎の程度を分ける分類法	なし
			体内診断	胆道膵管撮影時の消化管陰性造影	実薬高濃度群の低濃度群に対する優越性の検証を目的とした。投与前後の2枚の撮像フィルムを「相対的」に評価したため、低濃度群においても一定の陰性造影効果があると評価された。（追加試験では、投与前（無投与時）の撮像フィルムと投与後の撮像フィルムについて、それぞれ「絶対的」な評価に変更）	なし
⑦-2	主要評価項目の選択が不適切	臨床試験開始時の対象疾患の周辺状況を加味して、再度、前相までに使用した評価項目が適切であるか、再度、十分に検討する必要がある。	精神・神経用	レビー小体型認知症	適切な疾患教育・介護指導、あるいは介護者の自己学習により疾患理解度が向上し、介護方法が改善された可能性が高く、それに伴って患者の不安が軽減し、結果として、NPI-2*を用いた有効性評価を困難にしたものと推察された。（追加試験では主要評価項目をCIBIC-plusに変更） *NPI-2（Neuropsychiatric Inventory-2）：精神症状・行動障害検査であるNPI-12のうち、認知機能変動及び幻覚（幻視）の2項目の合計点数	なし

⑦-3	主観に関わる主要評価項目における評価方法の統一が不十分	主観による評価項目とする場合、例えば、患者日誌等を用いる際は、評価方法に関するトレーニング、記載に関する基準、また、小児の場合、記載者等、十分に検討し、評価方法が統一となるように工夫する必要がある。	アレルギー	アレルギー性鼻炎（小児）	临床上想定し難い多くのくしゃみ発作及び擤鼻回数を記録した被験者が散見されたことから、評価方法の統一化が適切に行われていなかった可能性があった。（追加試験では、鼻症状の評価方法を統一するため、評価方法に関するトレーニング方法を見直し、鼻アレルギー日誌のまとめ書きを防止すること、並びに被験者の鼻症状を適時把握することを目的に、電子患者日誌システムが用いられた。）	②-1
⑦-4	主要評価項目の評価時期が不適切	主要評価項目の評価時期について、臨床症状の変動時期、薬剤の薬物動態を踏まえて、適切に設定されているか、十分に検討する必要がある。	アレルギー	アレルギー性鼻炎	花粉曝露室への入室中はプラセボ及び実薬ともに鼻閉スコアが上昇しており、花粉曝露終了後（花粉曝露室退室後）は速やかに花粉曝露前の水準に戻っていたことから、花粉曝露室退室後の検査ではなく入室中の鼻症状であれば評価可能であることが示唆された。	②-1
			アレルギー	アレルギー性鼻炎（小児）	患者日記は、就寝前に1日の鼻症状を午前及び午後に分けて記録となっていたが、実薬1日1回投与群では血中濃度のトラフ時に症状の評価となり、小児では成人と比較して半減期が短いことも勘案すれば、症状の改善が明確に評価されなかった可能性があった。 注）本品は、患者日記を原則として小児である患者本人に記録させたため、正確な評価が困難であった可能性も考えられた。	④-1
			口腔乾燥症状改善	シェーグレン症候群患者の口腔乾燥症状	最終評価時は、唾液検査を実施するために治験薬を服用せずに来院し、最終服薬から時間が経過した状態で口腔乾燥重症度 VAS が測定された。実薬は、対照薬より最高血中濃度に到達するまでの時間が早く（実薬 1.3 時間、対照薬 2.5 時間）、速やかな唾液分泌促進作用を示すものの、消失半減期が短いため（実薬 1.4 時間、対照薬 4.3 時間）、投与から口腔乾燥重症度 VAS 測定までの時間が延長した影響がより強く現れた可能性があると考えられた。	なし

⑦-5	主要評価項目の評価期間が不十分	病態の悪化、再発・再燃等の好発時期等に関する情報を十分に精査し、主要評価項目の評価期間が適切であるか、十分に検討する必要がある。	皮膚疾患用	掌蹠膿疱症	正常表皮のターンオーバーが約 1.5 ヶ月とされていることを踏まえ、膿疱・小水疱を改善し、既に形成されている炎症性角化局面に存在する紅斑や角化・鱗屑までを改善するには、少なくとも 4 週間以上の期間が必要であり、4 週間の評価期間では不十分であった。	なし
			精神・神経用	パニック障害	本臨床試験の再燃及び海外臨床試験の成績より、パニック障害における再燃時の症状は、短期間で投与前の状態まで悪化することは少ないことが示唆された。	なし
			免疫抑制	重症筋無力症	ステロイド投与量は、実薬群で連続的に減少し、プラセボ群では症状の維持が困難となるため増加すると想定し、遅くとも投与開始 16 週後（投与終了 12 週前）までには、ステロイド投与量が安定すると予測していた。しかし、本試験におけるステロイド平均投与量はプラセボ群及び実薬群ともに、投与開始 20 週後まで経時的に減少し、プラセボ群において増加傾向が認められたのは投与開始 20 週後以降であり、試験計画時の予測と異なっていた。	なし
⑦-6	評価方法が定まっていない段階での検証的試験の実施	主要評価項目の評価方法について、検証的試験開始前までに決定することは重要である。	精神・神経用	球脊髄性筋萎縮症	本臨床試験開始時点において、主要評価項目に関する検討内容が十分であったとは言い難かった。	なし

⑦-7	主要評価項目の選択が不適切	主要評価項目の選定においては、計画時点で検討されている評価指標の情報等も考慮して、検討する必要がある場合がある。	歯科	歯周炎による歯槽骨の欠損	米国歯周病学会の Consensus report を参考に、「歯槽骨レベルの改善」及び「臨床的アタッチメントの獲得量」を主要評価項目に設定したが、後者では統計的な有意差は認められなかった。 しかし、その後、日本で発出された歯周組織治療用細胞シートに係る「次世代医療機器評価指標の公表について」では、歯周組織破壊を伴う歯周疾患（歯周炎）等の治療に際してフラップ手術単独でも上皮性付着を含む臨床的アタッチメントの獲得が認められることから、歯周組織再生の有効性評価は、「歯槽骨レベルの改善」を主要な評価項目にすることが推奨されたこと、類似の医療機器の臨床試験では「新生歯槽骨増加量」で有効性を評価していた。	なし
-----	---------------	--	----	--------------	---	----

ICH E8 「臨床試験の一般指針」

3.2.2.4 反応変数

反応変数は、試験開始前に規定され、観察及び定量化の方法を具体的に示すものでなければならない。可能かつ適切である限り、客観的な観察方法が用いられるべきである（ICH E9 ガイドライン（別表）参照）。

ガイドラインには、上記の通り、反応変数は、観察及び定量化の方法を具体的に示すものでなければならないと記載されているが、上記の一覧では、評価方法・判定基準の統一が不十分であった例が多くみられている。ガイドラインに明確な記載がないことから、一覧で具体的な事例を示すことは、ガイドラインを活用の助けとなると考える。

⑧ 統計手法

	項目	留意点	具体的な事例			
			薬効分類	適応	ネガティブな結果となった原因	関連項目
⑧-1	非劣性の評価方法が不適切	解析計画を検討する際、極端な結果（例：有効率 100%、発生率 0%等）となる可能性も考慮して、非劣性が評価できる方法を検討する必要がある。（正確な信頼区間を算出できる手法の選択等）	体内診断薬	腹部のコンピューター断層撮影における造影	事前に造影効果の有効率の群間差の 95%CI に基づいて実薬群の対照群に対する非劣性を評価する計画とされていたが、有効率は両群ともに 100%であり、事前に計画された方法での造影効果の有効率の群間差の 95%CI (Wald CI) は算出不能となった。両群の造影効果の有効率が 100%であった場合でも群間差の 95%CI を算出して非劣性が評価できる計画とすべきであった。	なし
⑧-2	評価対象集団が不適切	複数用量の実薬を併合した群で有効性を検証する計画とする場合、併合群に含まれる各群のうち、有効性に影響を与える群が含まれていないか十分に検討する必要がある。	循環器系	肺動脈性肺高血圧症	実薬併合群（低用量、中用量、高用量）とプラセボ群を比較した臨床試験で低用量群において改善が想定よりも小さくなり、低用量群の結果が併合群とプラセボ群の比較に影響し、併合群とプラセボ群の間に統計的に有意な差が認められなかったものと考えられた。 注) 本試験の場合、プラセボ群と比較する実薬群の用量の設定に関する検討が不十分であったことも影響していると考えられる。	なし
		ICH E9 ガイドラインの記述内容に示されているように、ランダム化が行われた被験者を主たる解析対象集団から、有効性評価に偏りを生じさせる危険性をもたらすことなく、除外できる状況は限られていることを踏まえ、不適切に、有効性解析対象集団から除外する計	抗ウイルス	新型又は再興型インフルエンザウイルス感染症	治験に組み入れられた 762 例のうち 122 例を除いた PPS 集団を主たる解析対象集団と設定、その元となる FAS 集団は、組入れ被験者の 1 割近くが別途開催された症例検討会において「対象外疾患」と判断され解析対象集団から除外されていた。 PMDA は、有効性の主たる解析対象集団は、ITT の原則に基づき FAS と設定することがより適切であったと判断した。また、	なし

		画になっていないか、検討する必要がある。 注) ICH E9 (R1) で議論されている estimand の概念を踏まえて、試験の目的、推定したい治療効果等を考慮して被験者、データの取扱いを予め決定しておくことが必要であることも示している。			症例検討会において「対象外疾患」とみなされ、解析対象から除外された症例は、RT-PCR 陰性例が 43 例、合併症例（疑いを含む）が 30 例、RT-PCR 陰性かつ合併症例が 1 例であった。RT-PCR 陰性例について、本臨床試験では日常診療と同様にインフルエンザキットにより陽性と診断された症例がランダム化されており、RT-PCR 陰性例を有効性解析対象集団からは除外すべきではなく、当該症例を含めた解析対象における解析結果を基に、有効性を評価することが適切と判断された。また、合併症例（疑いを含む）について、個々の症例の採否により解析対象集団を構成することは不適切であり、RT-PCR 陰性例以外の「対象外疾患」として除外された 31 症例は、治療の結果に依存するようなランダム化以後の主観的な判断に基づき除外の採否が決定されており、ITT の原則に従い、有効性解析対象集団からは除外すべきではないと考えた。	
⑧-3	主要解析における順位付けの検討が不十分	有効性が示唆されているものの、有害事象により中止した被験者における取扱いについても、事前に検討しておく必要がある場合がある。	精神・神経用薬	全身型重症筋無力症	レスキュー治療の実施基準への該当の有無にかかわらず、レスキュー治療を受けた患者とすべての中止例を併せてイベント発現までの期間が短い患者から最悪順位を付けたことは、重症筋無力症の症状は改善しているものの、有害事象により中止した患者の評価にバイアスを生じている可能性があった。 重症筋無力症の症状は改善しているものの、有害事象により中止した患者を保守的に取り扱ったために、実薬群の有効性が過小評価された可能性が高いと考えられた。	なし

⑧-4	欠測値の取扱いが不適切	ICH E9 ガイドラインの記述内容に示されているように、欠測値に対処する方法により解析結果がどの程度変わり易いかも踏まえて、欠測値の取扱いは十分に検討する必要がある。欠測値が生じにくいデザインになっているか検討すること、また、想定される欠測値（やむを得ず生じる）に関する事前の検討も必要である。	精神・神経用薬	線維筋痛症に伴う疼痛	欠測値を他時点の観測値等で補完をせずに、観測された全時点のデータを用いて変数の経時プロファイルを評価する手法である Mixed-effects Models for Repeated Measures (MMRM) 解析を主解析とした。BPI 疼痛重症度の変化量が小さい効果不十分・悪化による中止例がプラセボ群でより多く認められたため、MMRM 解析では、実薬群と比較してプラセボ群の投与後 14 週時点の変化量が改善方向へ過大に推定された可能性が考えられた。	なし
			精神・神経用薬	社会不安障害	投与開始 1 週以内の中止例は実薬の低用量群が多かった。低用量群は、共分散分析 (LOCF) において、薬剤の効果が十分に発揮されていない時期の値で欠測値が補完されたため、実薬の有効性が過小評価された可能性がある。なお、当該期間において、有害事象による中止例が実薬低用量群が多かったことは偶発的なものであると考えられた（高用量群は低用量から開始されていたが、高用量群での中止例は低用量群の投与量の際に有害事象が認められ中止していたため）。	なし

ICH E9 「臨床試験のための統計的原則」

5.2.1 最大の解析対象集団（一部抜粋）

ランダム化が行われた被験者を最大の解析対象集団から除外することになる状況は限られている。それらには、主要な登録基準を満たしていない場合（適格基準違反）、試験治療を一回も受けていない場合、ランダム化後のデータがない場合などがある。そのような除外については常に理由を示すべきである。登録基準を満たしていない被験者は、以下の条件下でのみ偏りを導入する可能性なく除外できるであろう。

- ・ 登録基準はランダム化以前に評価されている
- ・ 除外の対象となる適格基準違反の発見は完全に客観的になされる
- ・ すべての被験者が適格基準違反について同様の綿密さで調べられている（非盲検試験においてはこの保証は困難であり、二重盲検試験であっても割付を明らかにした後では難しい。このことは盲検下レビューの重要性を強調している。）
- ・ 特定の登録基準違反が発見された場合、それに関するすべての違反が除外される

上記の一覧においては、上記のガイドラインの記載を踏まえて、計画を立案すれば、避けることができた可能性がある事例も見られている。（⑧-2、抗ウイルス薬）

5.3 欠測値と外れ値

欠測値は、臨床試験において偏りを起こし得る代表的な原因である。したがって、データ収集及びデータマネジメントに関しては、治験実施計画書が要求する事項について、すべてを満たすようあらゆる努力をすべきである。しかし、実際には、ほとんど常に欠測値がいくつか存在するであろう。そうであったとしても、欠測値を扱う方法が適切で、特にそれらの方法が治験実施計画書の解析計画中に事前に定められている場合には、試験は妥当であるとみなすことができる。この観点から盲検下レビューの間に統計解析計画書を見直すことにより、欠測値を扱う方法が洗練されるであろう。残念ながら、欠測値に対処する方法で、普遍的に適用可能と薦められる方法はない。欠測値に対処する方法により解析結果がどの程度変わり易いかを、欠測の数が多い場合には特に、検討すべきである。

ガイドラインには、上記の通り、欠測値の取扱いに関する記載があるが、一覧で具体的な事例を示すことは、ガイドラインを活用の助けとなると考える。

⑨ その他

	項目	留意点	具体的な事例			関連項目
			薬効分類	適応	ネガティブな結果となった原因	
⑨-1	自然環境の影響を受ける試験の限界	花粉の飛散等の自然環境の影響により、試験間、試験内又は被験者間において環境条件をそろえることができず、当該内容が有効性の評価に影響する場合、抗原誘発試験等、自然環境の影響を受ける状況下の試験以外で実施することを考慮する必要がある。	アレルギー	アレルギー性結膜炎	本試験は、環境及び試験条件を整え、環境要因の排除を最大限に試みたものの、花粉飛散量の変動等の影響が依然として存在し、自然環境の影響を受ける状況下の試験において、プラセボに対する実薬の優越性を検証することは困難とされた。	なし
			アレルギー	アレルギー性結膜炎	1地点で測定した花粉飛散量を用いたスローブ解析では、花粉飛散量の地域差、被験者の生活範囲の多様性などにより、実際の被験者の曝露花粉量は測定地点の花粉飛散量と乖離していたため、当初計画したスローブ解析でプラセボに対する実薬の優越性の検証に至らなかった。	なし

以上