Rare disease の治療効果の推測法

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

2021 年度　タスクフォース 3
2022 年度　継続タスクフォース 3

Ver 1.0
2022 年 12 月

製薬協

目次

1. はじめに

1.1 日米欧の希少疾病用医薬品指定制度

希少疾患（rare disease）とは，患者数が極めて少ない疾患を指す。希少疾患の中には重篤で予後不良な難病が多く存在するが，その多くは十分な治療法が確立しておらず，新たな治療法の必要性が高い。

患者数が少ない希少疾病に対する医薬品においても，製造販売承認取得のためには臨床試験等によりその有効性・安全性を示す必要がある。しかし，患者数が少ないために通常の臨床試験の方法の適用が難しい場合があること，市場性が小さいために投資回収が容易でないこと等から，開発が躊躇される傾向にあった。

希少疾患の治療を目的とした医薬品を希少疾病用医薬品に指定し，開発を支援する制度が設けられている。日本では，1993 年より制度が開始され，患者数が少ない，或いは指定難病の疾患に対する，医療上特にその必要性が高い，かつ開発の可能性が高い医薬品に対して，指定を受けることができる[1]。米国では 1983 年[2]，欧州では 2000 年[3]から同様の制度が設けられているが，その違いについて表 1-1-1 に示す。希少疾病用医薬品に該当する患者数は，日本では患者数 5 万人未満（人口に占める割合は 0.04%未満），米国では患者数 20 万人未満（人口に占める割合は 0.06%未満），欧州では人口 1 万人当たりの患者数が 5 人未満（人口に占める割合は 0.05%未満）であり，多少異なる。また，日本のみ「開発の可能性」が指定要件に含まれており，対象疾病に対して当該医薬品等を使用する理論的根拠があるとともに，その開発に係る計画が妥当であると認められることも求められる。開発支援の内容としては，日本では助成金交付，優先対面助言，優先審査，再審査期間の延長，税額控除などの優遇措置を受けられ[4]，米国及び欧州においても表 1-1-1 に示す優遇措置が受けられる。

さらに日本では，希少疾病用医薬品を含む，重篤で治療法のない疾患に対する革新的な新薬を早期に実用化するため，2015 年には先駆け審査指定制度が開始され，2020 年には先駆的医薬品等の指定制度として法制化された。米国のブレークスルーセラピー指定（2013 年開始），及び欧州の PRIME 制度（2016 年開始）も同様の制度である。その他，日本では，医薬品条件付き早期承認制度や新薬創出・適応外薬解消等促進加算制度があり，希少疾病用医薬品でも活用することができる。

日米欧で細部に違いはあるものの，薬事上の特別措置による希少疾病用医薬品の開発促進策が講ぜられている。

表 1-1-1：日米欧の希少疾病用医薬品指定制度

	日本	米国	欧州
開始	1993 年	1983 年	2000 年
指定の条件	・国内の対象患者数 5 万人未満または指定難病に指定されている ・医療上特にその必要性が高い ・開発の可能性が高い	・国内の対象患者数 20 万人未満 ・当該医薬品の開発，及び製造に要する費用の十分な回収ができない医薬品	・欧州内での対象患者数が 1 万人当たり 5 人未満 ・生命を脅かす，非常に重篤な疾患の治療，予防または診断を目的とする ・医療上特にその必要性が高い
優遇措置	・治験相談，及び審査手数料の減額 ・試験研究費の助成金交付 ・試験研究費に対する税制措置 ・薬価への加算 ・優先的な対面助言，及び審査	・審査手数料の減額 ・試験研究費への助成金交付 ・試験研究費に対する税制措置	・審査手数料等の免除または減額 ・試験研究費への助成金交付 ・試験計画作成の補助 ・中央審査方式を利用できる
市場独占期間	10 年 （再審査期間として）	7 年	10 年

参考文献

[1] 厚生労働省，“希少疾病用医薬品等の指定に関する取扱いについて”，薬生薬審発 0831 第 7 号，薬生機審発 0831 第 7 号（令和 2 年 8 月 31 日）

[2] Food and Drug Administration. “Developing Products for Rare Diseases & Conditions.” https://www.fda.gov/industry/developing-products-rare-diseases-conditions

[3] European Medicines Agency. “Orphan designation: Overview.” https://www.ema.europa.eu/en/human-regulatory/overview/orphan-designation-overview

[4] 厚生労働省，“希少疾病用医薬品・希少疾病用医療機器・希少疾病用再生医療等製品の指定制度の概要”，https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000068484.html

1.2　　最近の希少疾病用医薬品開発が置かれている環境

希少疾患は 6,000 から 8,000 種類あることが知られている。これらのうち 70%以上は遺伝性疾患であり，小児から発症することが多い。希少疾患それぞれの患者数は少ないが，希少疾患に罹患している患者総数は，世界人口の 3.5~5.9%に相当する 2 億 6,300 万人~4 億 4,600 万人と推定されており，大規模である [1, 2]。

2016 年から 2020 年までの過去 5 年間に承認された新有効成分含有医薬品のうち，希少疾病用医薬品が占める割合は日米欧でそれぞれ 30.6%，49.6%，36.4%であった[3, 4, 5]。2020 年にまとめられた報告書[6]では，希少疾病用医薬品の承認件数は，今後日本を含む全世界で増加していくと予想されている。その理由として，遺伝子技術をはじめとした技術革新が進んでいること，また，高度かつ有用性の高い個別化医療も実現してきていることから，希少難病に対する革新的な医薬品の創出が可能となってきていることを挙げている。

希少難病に対する革新的な医薬品という意味での新たなモダリティーに着目すると，その一つとして核酸医薬品が挙げられる。核酸医薬品は，遺伝子の構成成分である核酸と類似した構造を持ち，特定の遺伝子配列を標的として，その遺伝子から作られるタンパク質の産生を調節することで作用する。核酸医薬品により従来の低分子医薬品では難しかった疾患の治療が可能になると期待される[7]。例えば，ビルトラルセンは遺伝性筋疾患であるデュシェンヌ型筋ジストロフィー（筋ジストロフィーは指定難病）の治療薬として 2020 年に日本，及び米国で承認された核酸医薬品であり，日米欧で希少疾病用医薬品の指定を受けている。また，希少難病に対する新たな治療法としては，再生医療等製品による治療も挙げられ，2021 年 6 月時点で日本では 25 品目が希少疾病用再生医療等製品指定を受けている[8]。希少がんの領域では，近年の革新的な分子生物学的解析法の進歩により，これまでの発生臓器・病理形態分類とは別に，ゲノム情報によるがんの分子分類が可能になりつつある。一部のがんでは単一の遺伝子変異ががん発生の直接の原因（driver）であることが明らかとなってきた[9]。エヌトレクチニブは，*neurotrophic tyrosine receptor kinase (NTRK)* 融合遺伝子陽性の進行・再発の固形がんに対する承認を 2019 年に日本，及び米国で取得しており，希少な遺伝子変異を有する固形がんに対し，がん種を問わず使用が認められている。

一方，薬効評価に目を向けてみると，以前より挙げられていた課題として，希少疾病用医薬品を対象とした検証的な臨床試験では対象患者が少ないことにより仮説検定に対して十分な検出力を確保する症例数で試験を実施できない，疾患の重篤性によりプラセボ対照を置くことは倫理的に不可能である等が挙げられていた。しかし，近年公刊されたガイドラインや報告書は，この課題の解決に一定の示唆を与えている。

米国においては，2019 年 1 月に米国食品医薬品局（FDA）から Rare Diseases: Common Issues in Drug Development [10]がドラフトガイダンスとして公刊された。ガイダンスの目的として希少疾患領域におけるより効率的かつ効果的な医薬品開発プログラムの実施の支援が挙げられている。同じく FDA から同年 9 月には Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products [11]が公刊された。このガイダンスは主としてベイズ流アプローチを用いた

革新的な試験デザインに焦点が当てられている。日本においては，医薬品医療機器総合機構（PMDA）の外部機関である科学委員会から，2017 年 11 月に希少がんの臨床開発を促進するための課題と提言に関する報告書[9]が作成されており，希少がんの臨床試験デザインとしてベイズ流アプローチ等を用いた先進的デザインの利用について触れられている。また，日本医療研究開発機構のタスクフォースからも，試験開始前に日本の規制当局と協議すべきベイズ流アプローチの利用に関連する規制上の問題点を整理しつつ，その利用が望まれる臨床試験を挙げた成果物が作成されている[12]。

上記のような状況を鑑み，本タスクフォースは希少疾病用医薬品の治療効果の推測に関する統計手法，及び最近の適用事例について調査し，「Rare disease の治療効果の推測法」として報告書にまとめた。報告書には主に統計担当者を対象に統計手法に対する数理的解説を含め，希少疾病用医薬開発に関連するガイダンスや適用事例の紹介は医薬品開発に携わる担当者を広く対象読者としている。本報告書の構成は 1.3 節に記載する。

参考文献

[1] Nguengang Wakap, Stéphanie, et al. "Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database." European Journal of Human Genetics 28.2 (2020): 165-173.

[2] Dawkins, Hugh JS, et al. "Progress in rare diseases research 2010–2016: an IRDiRC perspective." Clinical and translational science 11.1 (2018): 11.

[3] 日本製薬工業協会. “承認取得品目データベース.” 承認取得品目の臨床データパッケージの調査・分析 | 医薬品評価委員会の成果物 一覧 | 日本製薬工業協会（jpma.or.jp）

[4] Food and Drug Administration. “New Drugs at FDA: CDER’s New Molecular Entities and New Therapeutic Biological Products”. New Drugs at FDA: CDER’s New Molecular Entities and New Therapeutic Biological Products | FDA

[5] The European Medicines Agency. “Annual reports and work programmes” Annual reports and work programmes | European Medicines Agency (europa.eu)

[6] 医薬産業政策研究所. “希少疾病用医薬品（Orphan drug）の開発動向 -FDA で承認された Orphan drug の日本での開発状況の分析-.” 医薬産業政策研究所 (2020).

[7] 日本医療研究開発機構. “デュシェンヌ型筋ジストロフィー治療薬（NS-065／NCNP-01，ビルトラルセン）の製造販売承認について” デュシェンヌ型筋ジストロフィー治療薬（NS-065／NCNP-01，ビルトラルセン）の製造販売承認について | 国立研究開発法人日本医療研究開発機構（amed.go.jp）

[8] 厚生労働省. “希少疾病用医薬品・希少疾病用医療機器・希少疾病用再生医療等製品の指定制度の概要，希少疾病用再生医療等製品に指定された品目一覧（令和 3 年 6 月 11 日時点）”

20210611 希少疾病用再生医療等製品指定品目一覧（mhlw.go.jp）
[9] 医薬品医療機器総合機構. “科学委員会 希少がん対策専門部会, 希少がんの臨床開発を促進するための課題と提言 2017 ― アカデミア及びレギュラトリーサイエンスの視点から ―.” 医薬品医療機器総合機構 (2017).
[10] Food and Drug Administration. "Rare Diseases: Common Issues in Drug Development - Guidance for Industry (Draft)." (2019).
[11] Food and Drug Administration. “Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products (Draft).” (2019).
[12] Akihiro Hirakawa, et al. “Regulatory issues and the potential use of Bayesian approaches for early drug approval systems in Japan.” Pharmaceutical Statistics (2022).

1.3　本報告書の構成

本報告書の構成について示す。2 章では FDA, Guidance for Industry, Rare Diseases: Common Issues in Drug Development を始めとする希少疾病用医薬品の治療効果の推測に関連する国内外のガイダンスを紹介する。3 章では，希少疾病用医薬品開発において主に検証段階の試験で用いられる，又は今後新たに用いられる可能性がある試験デザイン及び治療効果の推測に関する解析方法について解説し，関連する適用事例を示す。4 章では，本報告書のまとめを示す。なお，3 章で扱う内容が多岐に渡るため，以下に 3 章の各節の要約を示す。最後に本報告書の補遺に載せた項目を示す。

<3 章の各節の要約>

3.1 ヒストリカルコントロールの利用

同時対照の設定が困難な場合に用いられるヒストリカルコントロールに関して，Real World Data 等のヒストリカルコントロールのリソース，利用時の潜在的バイアスへの対処，ベイズ流階層モデルや Power prior などヒストリカルコントロール利用時の解析手法及び適用時の留意点を述べる。最後にヒストリカルコントロールを利用した事例を示す。

3.2 アダプティブデザイン

アダプティブデザインでは，群逐次計画などアダプテーションの具体例について述べ，用量選択に関するアダプテーションを伴うシームレス第 II/III 相試験の事例を解析手法とともに示す。

3.3 エンリッチメント戦略

エンリッチメント戦略では，大別される 3 種類のエンリッチメントを示し，中間解析による症例数再設定を伴う特定の患者層へのエンリッチメント戦略の事例を解析手法とともに示す。

3.4 マスタープロトコル

マスタープロトコルで扱われるアンブレラ試験，バスケット試験及びプラットフォーム試験について，それぞれの試験デザインの特徴を示し，マスタープロトコルの革新性を共通検査プラットフォームの利用などの実施基盤の観点及び試験デザインの観点から解説する。希少フラクションでがん種横断的に承認を取得した製剤の開発で用いられたバスケット試験の事例を示す。

3.5 小児開発における成人データの借用

小児開発において，FDA の承認審査時の照会で主要解析の結果を補足するためにベイズ流アプローチで成人データを借用した解析が求められ，対応した事例を示す。

3.6 試験デザインと解析手法についてのその他の議論

FDA Complex Innovative Design パイロットプログラムの中で実施された試験の事例を示す。次に，患者個人の治療効果推測に焦点をあてた N-of-1 試験デザイン及び複数の N-of-1 試験のメタ・アナリシスによる統合に基づく集団の治療効果推測について述べる。また，より少ない症例数でかつ治療効果推測の精度確保に関する最近の議論として，Complete N-of-1 試験デザイン，Probability of inconclusive 及び Probability monitoring

procedure を紹介する。

3.7 有意水準を両側 5%超で設定した試験

検証的試験において両側有意水準 5%（または片側有意水準 2.5%）よりも大きな有意水準を採用した事例を示す。

3.8 代替評価項目の活用

FDA の「医薬品承認の根拠となった代替評価項目一覧表[注)]」にも記載のある代替評価項目を使用して承認に至った事例を示す。

注）https://www.fda.gov/drugs/development-resources/table-surrogate-endpoints-were-basis-drug-approval-or-licensure

<補遺>

1. 群逐次法におけるファミリーワイズの第 1 種の過誤確率の制御
2. Lung-MAP 試験，及び ISPY-2 試験
3. N-of-1 デザインの症例数設定
4. Complete N-of-1 デザインの症例数設定

2. 関連するガイダンスの紹介

近年に発出された希少疾患の開発に役立つと考えられるガイダンスを 5 つ紹介する。これらの中でも 2.1 節で紹介する Rare Diseases: Common Issues in Drug Development は希少疾患の開発全般を対象としたガイダンスである。本章での紹介は，各ガイダンスの Introduction，Background，または他の部分の記載を要約したものである。

なお，特定の希少疾患についての臨床評価に関するガイドラインは，本タスクフォースでの検討には含めていない。

2.1 Rare Diseases: Common Issues in Drug Development

FDA から 2019 年に発出されたドラフトガイダンス“Rare Diseases: Common Issues in Drug Development”[1]は，希少疾病の治療または予防を目的とする医薬品及び生物学的製剤を開発する治験依頼者が，より効率的な医薬品開発プログラムを推進することの支援を目的としている。そして，以下の 8 点が重要であると述べている。

- 疾患の自然歴に関する十分な理解
- 疾患の病態生理，及び薬剤の作用機序に関する十分な理解
- 臨床試験による治療の安全性を裏付けるための毒性に関する考察
- 評価項目の選択または開発
- 安全性及び有効性を確立するためのエビデンス
- 医薬品開発中における製造上の留意事項
- 患者，介護者，及び支援者の開発プログラムへの参加
- 規制当局との対話

ガイダンスの序文には，医薬品の製造販売承認のための要件は，希少疾病に対しても一般的な疾患と同じであることが記載されている。そして，このガイダンスで議論されている点は，他の医薬品開発プログラムでも遭遇することがあるが，医学的，及び科学的知識，自然歴データ，医薬品開発経験が限られていることが多い希少疾病の状況下では，より対処が困難なことが多いと述べられている。

ガイダンスで紹介されている，いくつかの話題について，以下に紹介する。

<u>自然歴研究について</u>

希少疾患の自然歴は十分に理解されていないことが多く，前向きにデザインされた，治験実施計画書に基づく自然歴研究を医薬品開発計画の初期に開始することは非常に重要である。自然歴研究による疾患への理解は，以下の点で治験依頼者の助けとなりうる。

- 疾患の状態の範囲，及び重要な疾患サブタイプの特定を含め，疾患集団を定義することができる。これにより，疾患の進行が速い患者を選択し，臨床試験で用いる評価項目を開発することができる（予後エンリッチメント）。

- 臨床試験デザインにおける重要な要素である試験期間や症例登録基準を適切に設定できる。
- 感度及び特異度が高い評価項目を開発することができる。
- Proof-of-concept（POC），用量選択，潜在的なレスポンダーのスクリーニング（予測エンリッチメント），安全性上の懸念の早期発見，または有効性の裏付けとなるエビデンスに関して，有益な情報をもたらす可能性がある新規バイオマーカーを特定することができる。または既存バイオマーカーのバリデーションを行うことができる。

また，試験内で同時対照群を設定することが非現実的または非倫理的である場合などの特別な状況では，適切にデザインされた自然歴研究を介入試験の外部対照群として用いることができる。なお，エンリッチメント戦略の詳細及び適用事例については 3.3 節を参照されたい。

ヒストリカルコントロール群の利用について

医療上の必要性が満たされていない重篤な希少疾病に対しては，登録されたすべての患者が試験薬の投与を受けて，同時対照群（例：プラセボまたは標準治療）へのランダム化を伴わないような，ヒストリカルコントロール群の利用に関心が示されることが多い。しかし，同時対照群でないために生じる系統的な差を排除できないことが，ヒストリカルコントロール群を用いるデザインの大きな問題である。一般的にヒストリカルコントロールの利用は重篤な疾患の評価に限定され，以下の（1）から（3）を満たす必要がある。

（1）アンメットメディカルニーズが存在すること
（2）高い死亡率など，十分に立証された，客観的に測定・検証可能な疾患経過が存在すること
（3）予想される薬物効果が大きく，自明であり，介入と時間的に密接に関連していること

しかしながら，臨床経過の予測性が高く，客観的に検証可能なアウトカム指標を有する疾患であっても，ヒストリカルデータでは，知られていない，または記録されていない重要な共変量が存在する可能性に留意が必要である。なお，ヒストリカルコントロールの利用に関する詳細及び適用事例については 3.1 節を参照されたい。

2.2 Rare Diseases: Natural History Studies for Drug Development

FDA から 2019 年に発出されたドラフトガイダンス“Rare Diseases: Natural History Studies for Drug Development”[2]は，希少疾病用の医薬品，及び生物学的製剤の開発を支援するために用いられる自然歴研究のデザインと実施に関する情報提供を目的としたものである。

2.1 節で紹介した Rare Diseases: Common Issues in Drug Development[1]では，希少な疾患を対象とした医薬品開発において遭遇する共通の問題を検討しているが，本ガイダンスでは自然歴研究を主題として取り上げている。

このガイダンスでは，希少疾患を対象とした医薬品開発の全ての段階における自然歴研究の幅広い潜在的用途，様々な種類の自然歴研究の長所と短所，データ項目，及び研究計画，並びに

自然歴研究を実施するための枠組みについて述べられている。また，試験デザインを試験の目的と整合させて試験結果の解釈可能性を高めるための考慮事項や，自然歴研究における患者のデータ保護の問題についても論じられている。

2.3 承認申請等におけるレジストリの活用に関する基本的考え方

厚生労働省から 2021 年に発出された「承認申請等におけるレジストリの活用に関する基本的考え方」[3]は，希少疾病のような患者数等の限界から比較試験の実施が困難な場合に薬効を評価する方法として，レジストリデータの活用を促進する目的で作成された。

ガイダンスの適用範囲としては，主にレジストリデータを活用する場合が想定されているが，過去に行われた治験のプラセボ群等のデータを活用する場合においても，参照可能な部分はあるとされている。

このガイダンスにおいては，承認申請等にレジストリデータを活用する場合に，一般的に考慮すべき点として，以下の 4 点が挙げられている。

(1) 個人情報の保護に関する配慮，及び患者の同意
(2) 活用するレジストリデータの信頼性
(3) 活用するレジストリデータの適切性
(4) レジストリを構築する者（レジストリ保有者）との早期からの協議

また，臨床試験においてレジストリデータを外部対照等として承認申請等における有効性及び／または安全性の評価に活用する場合に考慮すべき内容として，以下の 5 点が挙げられている。

(1) レジストリの患者集団
(2) 評価項目
(3) 評価期間
(4) 統計手法
(5) 自然歴の観察研究のタイプ（前向き，後向き）

なお，上記の点は全ての状況で必要不可欠な点を示しているわけでないとされており，活用に当たっては PMDA が実施する対面助言を活用することが強く推奨されている。

2.4 Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products

FDA から 2020 年に発出されたガイダンス“Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products”[4]は，治験依頼者，及び申請者に対し，医薬品または生物学的製剤に関する複雑で革新的な試験デザイン（CID）についての，FDA との対話に関する指針を示す目的で作成された。

このガイダンスでは，医薬品及び生物学的製剤の開発や審査における新規の試験デザインの使用，モデリング&シミュレーションに関連する技術的課題に関して，治験依頼者が FDA からどの

ようにフィードバックを得ることができるか，並びに審査のために提出すべき定量的及び定性的情報について論じられている。

CID は複雑なアダプティブ，ベイズ流，及びその他の新規の臨床試験デザインを指すと考えられているが，革新的または新規と考えられるものは時間とともに変化する可能性があるため，CID の固定された定義はないとされている。そして，ガイダンスの中ではマスタープロトコル，Leveraging Data From Phase 2 to Phase 3，及び Sequential Multiple Assignment Randomized Trials（SMARTs）が CID として例示されている。なお，CID の詳細及び適用事例については 3.6.1 節を参照されたい

2.5 Master Protocols: Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologics

FDA から 2022 年に発出されたガイダンス“Master Protocols: Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologics”[5]は，成人及び小児の 2 つ以上のがん種に対して，及び／または 2 つ以上の試験薬を同時に評価することを目的とした臨床試験のデザインと実施に関して，推奨事項を提供する目的で作成された。なお，このガイダンスは，First in Human の臨床試験は対象としていない。

このガイダンスでは，マスタープロトコルの種類についての紹介がされた後に，以下の点が議論されている。なお，マスタープロトコルの詳細及び適用事例については 3.4 節を参照されたい。

- 試験デザインに関する留意点
- バイオマーカー開発における留意点
- 統計解析に関する留意点
- 安全性に関する留意点

参考文献

[1] Food and Drug Administration. "Rare Diseases: Common Issues in Drug Development - Guidance for Industry (Draft)." (2019).

[2] Food and Drug Administration. "Rare Diseases: Natural History Studies for Drug Development - Guidance for Industry (Draft)." (2019).

[3] 厚生労働省. “承認申請等におけるレジストリの活用に関する基本的考え方.” (2021).

[4] Food and Drug Administration. "Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products - Guidance for Industry." (2020).

[5] Food and Drug Administration. "Master Protocols: Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologics - Guidance for Industry." (2022).

3　　希少疾病用医薬品の検証的試験で用いられる試験デザイン，及びその事例

2 章で紹介した FDA の Rare Diseases: Common Issues in Drug Development [1]では，希少疾病を対象にした臨床試験を実施する際の留意点として，自然歴に関する十分な理解，外部対照の活用時の留意点，有効な患者層を特定するためのバイオマーカーの特定，及びエンリッチメント戦略について言及されていた。また希少疾病では登録可能な症例数が限られることから，CID を含むアダプティブデザイン，及び様々な疾患サブタイプを含めて 1 つの試験として実施するマスタープロトコルを用いた効率的な試験実施が必要な場合もあるだろう。また，希少疾病を対象にした臨床試験では有意水準の設定を通常よりも大きくすることや一般的に登録が難しいとされる小児を対象とした臨床試験であれば成人データの借用といったことも考えられる。

そこで本章では，希少疾病用医薬品の検証試験で用いられる治療効果の推測に関する統計手法について概説するとともに，上記のアプローチが実際に希少疾病を対象とした臨床試験で用いられた事例について，承認された事例を中心に紹介する。各事例では公開されている情報を元に，どういった背景で各アプローチが適応され，規制当局とどういった議論がされていたかをまとめた。一部，申請に至らなかった事例も含めたが，規制当局と合意の下，試験が計画・実行されたものである（ただし，マスタープロトコルの事例として挙げた DIAN-TU が実施するプラットフォーム試験については規制当局との議論は確認できておらず，したがってこれに当てはまらないことにご留意いただきたい。）。

なお，希少疾病を対象とした医薬品は数多くあるため，本タスクフォースでは網羅的な事例収集は実施しておらず，系統的なレビューも実施していない。本報告書の事例紹介は，論文情報や承認審査情報から，興味深いと考えられた事例を紹介することを意図している。

本章の内容が，今後希少疾病を対象とした臨床試験を実施する際に参考になれば幸いである。

参考文献

[1] Food and Drug Administration. "Rare Diseases: Natural History Studies for Drug Development - Guidance for Industry (Draft)." (2019).

3.1　　ヒストリカルコントロールの利用

医薬品の承認申請においては，同時対照群を設けた検証的なランダム化比較試験で，仮説検定を用いて有効性の検証を行うことが一般的であるが，この方法により仮説の検証を成功させるためには，想定効果サイズ (群間差 / 標準偏差) に応じた一定規模の試験が必要となる。しかしながら，症例登録のあらゆる努力を行っても，合理的な時間枠の中では，適切な規模での試験実施が不可能な場合がある[1, 2]。また，希少疾患には生命を脅かす疾患が多く，不可逆的な転帰を辿る疾患や，小児を対象とした臨床開発など，同時対照群の設定が倫理的な観点から難しい場合も存在する。このような同時対照群の設定が困難な場合においては，対照群もしくはその一部としてヒストリカルコントロールデータを利用し，新たな臨床試験で収集する治療群のデータと比較することが考えられる。ヒストリカルコントロールを利用することで，臨床試験に必要な費用の低減や期間の短縮が可能となり，臨床試験の実施可能性や効率性を高めることも期待できる。希少疾病やがん領域では，単群試験が実施されることが多く，外部対照群としてヒストリカルコントロールを使った結果や，Real World Data（RWD）の標準治療を外部対照群とした試験も受け入れられている[3]。

ヒストリカルコントロール群の利用に関しては，日本では「承認申請等におけるレジストリの活用に関する基本的考え方」[4]が 2021 年に発出され，米国では Rare Diseases: Common Issues in Drug Development）[5]が発出されている。さらに，関連するガイドラインとして，以前から ICH E10 ガイドライン「臨床試験における対照群の選択とそれに関連する諸問題」[6]，Guideline on Clinical Trials in Small Populations （EMA）[7]が存在する。また，規制当局から発出されたガイダンスではないが Pocock [8]による基準も良く知られている。

参考文献

[1] 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. "Small Clinical Trials による薬効評価の考え方."（2013）.

[2] 小宮山靖，日本製薬工業協会 医薬品評価委員会. "「医療イノベーション 5 か年戦略後の医薬品開発の姿」 -研究開発から製造販売後までの医療イノベーションへの挑戦-." 第 119 回医薬品評価委員会総会 講演要旨集（2012）； 58-59.

[3] Lim, Jessica, et al. "Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: review of methods and opportunities." Therapeutic innovation & regulatory science 52.5 (2018): 546-559.

[4] 厚生労働省. "承認申請等におけるレジストリの活用に関する基本的考え方", 厚生労働省 (2021).

[5] Food and Drug Administration. "Rare Diseases: Common Issues in Drug Development - Guidance for Industry (Draft)." (2019).

[6] European Medicines Agency. "Guideline on Clinical Trials in Small Populations." (2006).

[7] 厚生労働省. "臨床試験における対照群の選択とそれに関連する諸問題." (2001).

[8] Pocock, Stuart J. "The combination of randomized and historical controls in clinical trials." Journal of chronic diseases 29.3 (1976): 175-188.

3.1.1　主なヒストリカルコントロールのリソース

本節では，Ghadessi et al. (2020) [1]で示されている主なヒストリカルコントロールのリソースを紹介する。データの構造や品質はリソースによって様々であり，リソースの種類に応じた様々なバイアスや懸念点がある。そのため，臨床試験でヒストリカルコントロールを利用するにあたって，これらの点を十分に理解する必要がある。

3.1.1.1　Real World Data（RWD）

主な RWD として，カルテ（Medical chart），患者レジストリ（Patient registry），自然歴研究（Natural history study）が挙げられる。また，適応外使用の公表データや様々なソースから定期的に収集された患者の健康状態などのデータも RWD に含まれる。そのため，RWD を利用することで，通常の臨床試験では得られないような，より幅広い患者集団での情報を得られる可能性がある。一方で，RWD は承認申請資料等への二次利用を前提とせずに設計・運用されていることも多く，品質や適切性等の観点から承認申請等に利用可能か検討する必要がある。

3.1.1.1.1　カルテ

カルテは患者の病歴や臨床データ，その施設での医療情報などを網羅的に記録したものであり，他の施設での記録が含まれることもある。これには，医師，看護師，検査技師などの医療スタッフによって作成されたさまざまな医療メモや検査，診断，治療手順のデータが含まれる。カルテデータの二次利用では，欠測データの多さやデータの不正確さ，フリーテキストで記録されるという点から，データの構造や品質の観点で扱いが難しいといえる[1]。これは，カルテが紙であるか電子であるかを問わず同様である。

カルテデータに含まれている薬剤使用等の曝露情報やアウトカムの発生に関する情報が真実と異なる場合，真の曝露やアウトカムと異なる分類がされることによる誤分類バイアスが生じることがある。また，患者背景が異なる等の理由により，臨床試験の患者における治療効果は日常診療の患者より大きくなる傾向があることがいくつかの研究で示されている[2, 3]。そのため，カルテデータをヒストリカルコントロールとして推定した治療効果には患者選択バイアスが含まれている可能性がある。

3.1.1.1.2　患者レジストリ

患者レジストリでは，特定の疾患や状態，曝露によって定義された集団での特定のアウトカムに関するデータを収集する。Duke 大学と FDA の協働で設立された The Clinical Trials Transformation Initiative（CTTI）では，承認申請を目的としたレジストリ組み込み型臨床試験の実施に適したレジストリの評価とデザインに関する推奨事項をまとめている[4]。また，PMDA では，医薬品，医療機器，及び再生医療等製品の薬事申請へのレジストリ活用に関して，個々の品目の活用目的に即して，レジストリ使用の妥当性及びレジストリデータの信頼性についての相談枠（レジス

トリ使用計画相談及びレジストリ信頼性調査相談）を 2019 年度に新設した[5]。

3.1.1.1.3 自然歴研究

自然歴研究では，疾患の自然経過を追跡し，疾患の進行や，未治療におけるアウトカムに関連する人口統計学的，遺伝的，環境的要因やその他の変数を特定する。自然歴研究の選択基準は，疾患の不均一性とアウトカムに対する共変量の影響を特徴付けられるようにするため，広くすべきである。ただし，自然歴研究では，現在の標準治療を受けている患者が含まれることが多く，これにより疾患の進行が抑制される可能性がある。利用可能な既存の自然歴研究がない場合，介入試験の開始前，かつ非臨床を含む医薬品開発の初期段階で自然歴研究を始めることが推奨される。しかし，多くの希少疾患は進行が急速あるいは生命を脅かすものである。そのため，そのような状況下で介入試験の前に非介入試験として自然歴研究を実施することが介入試験の開始を遅らせるような場合は，倫理的な妥当性が問われる。

3.1.1.2 完了した臨床試験

同じ作用機序または同じ薬剤の完了した臨床試験のデータは，管理された環境から得られた高品質なデータであるため，ヒストリカルコントロールのリソースとして優れている。完了した臨床試験の対照（プラセボまたは標準治療）群のデータをヒストリカルコントロールの候補として考えることができる。

参考文献

[1] Ghadessi, Mercedeh, et al. "A roadmap to using historical controls in clinical trials–by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG)." Orphanet Journal of Rare Diseases 15.1 (2020): 1-19.

[2] Clarke, Mike, and Kirsty Loudon. “Effects on patients of their healthcare practitioner's or institution's participation in clinical trials: a systematic review.” Trials 12.16. (2011).

[3] Smith, Malcolm A, and Steven Joffe. “Will my child do better if she enrolls in a clinical trial?.” Cancer 124.20 (2018): 3965-3968.

[4] The Clinical Trials Transformation Initiative. "CTTI Recommendations: Registry Trials." https://ctti-clinicaltrials.org/wp-content/uploads/2021/06/CTTI_Registry_Trials_Recs.pdf.

[5] 西岡絹恵ら. "PMDA におけるリアルワールドデータの活用にかかわる取り組み—新規相談枠の紹介—." レギュラトリーサイエンス学会誌 9.3 (2019): 197-204.

3.1.2　ヒストリカルコントロールの利用可能性と潜在的バイアスへの対処

ICH E10 ガイドライン[1]では，対照群の種類の 1 つとして外部対照群が挙げられており，詳細な検討がなされている。ガイドラインでは，外部対照群の使用は，治療効果が劇的であり，疾患の通常の経過が十分に予測可能である場合に限定され，外部対照を採用するのは，評価項目に対するベースラインや治療変数の影響の特徴が十分にわかっている場合に限ることが述べられている。

Rare Disease: Natural History Studies for Drug Development（FDA）[2]では希少疾患に対する薬剤開発における，自然歴研究の主な役割が述べられており，自然歴研究を使った外部対照群について言及されている。また，Framework for FDA's Real-World Evidence Program（FDA）[3]では，医薬品，及び生物学的製剤の効能承認の意思決定支援，及び承認後の安全性調査の要件設定支援のための Real World Evidence（RWE）創出にあたって，Real World Data（RWD）の潜在的利用を評価する RWE Program のフレームワークを提示しており，外部対照の利用可能性が言及されている。ここでは，医療習慣の潜在的な違い等により，比較可能な集団を確実に選択することが困難であるものの，傾向スコアのような統計手法を用いることで，外部対照データであるヒストリカルコントロールとの比較可能性の改善が期待できることが述べられている。

臨床試験でのヒストリカルコントロールの役割として，単群試験の外部対照群としたり，事前に定義したアルゴリズムやマッチング等の統計手法を用いて，二重盲検試験の対照群と併合することが考えられる[4]。しかしながら，ヒストリカルコントロールはランダム化や盲検化がされた試験の同時対照のデータではないため，以下のような潜在的バイアスの存在が懸念される。ヒストリカルコントロールの潜在的バイアスを低減させるための方策として，利用するヒストリカルコントロールの基準を設けることや解析手法による対処が考えられ，これらは試験実施前に決定しておく。利用するヒストリカルコントロールの基準や解析手法については，レジストリ使用計画相談を活用して，試験実施前に PMDA と協議することも可能である。

<u>選択バイアス</u>

ヒストリカルコントロールと新たな試験データでは，標準的なベースライン特性だけでなく，医療環境，併用療法，標準治療，心理的影響など様々な項目が異なることが想定され，これらが群間の体系的な差異となる。

<u>評価バイアス</u>

盲検化がなされていないため，主観的な評価項目の評価に影響を与える。

Pocock（1976）[5]は，新規試験の対照群の一部として利用可能な既存のヒストリカルコントロールの条件（Pocock の基準）として，以下の 6 つを挙げている[6]。

<u>Pocock の基準</u>

1. 詳細に定義された，新規試験の対照治療群と同じ治療を受けている。

2. 新規試験と同じ適格基準を用いて最近実施された臨床試験である。
3. 治療の評価方法が新規試験と同じである。
4. 重要な患者背景の分布が新規試験と同様である。
5. 新規試験とほぼ同じ組織で実施されている。
6. 新規試験との結果の違いを生むと予想される他の要因が存在しない。例えば，新規試験において症例登録スピードが予想よりも上がらないとき，適格基準をみたすぎりぎりの症例が登録されることも起こりうる。そういった場合，新規試験と既存試験で共通の適格基準を用いたとしても登録患者の背景が二つの試験間で異なってくる可能性がある。

Pocock（1976）[5]は，以上の 6 条件をすべて満たせばヒストリカルコントロールデータを新規試験の一部として安全に利用できるが，ひとつでも条件が満たされないときは，治療法の比較に大きなバイアスが生じる可能性を否定できないと指摘している。しかし，全ての条件を満たすヒストリカルコントロールデータが存在することは現実的にほとんどない。一部の条件が満たされていない状況で，バイアスを低減させながらヒストリカルデータの情報を利用する方法として，ヒストリカルデータと新規試験データの類似度に応じてヒストリカルデータから借用する情報量の大きさを制御する方法（階層モデルや power prior を用いた方法）やヒストリカルデータと新規試験データ患者の背景因子（ベースライン共変量）の不均一性を調整する方法（傾向スコアを用いた方法）が考えられる。3.1.3 節ではヒストリカルコントロールを利用する解析手法として一般的に提案されている，階層モデル，power prior，及び傾向スコア等のマッチングに基づくアプローチを紹介する。また，3.1.4 節ではこれらの解析手法を利用する際の留意点を述べる。

参考文献

[1] 厚生労働省. "臨床試験における対照群の選択とそれに関連する諸問題." (2001).
[2] Food and Drug Administraion. "Rare Diseases: Natural History Studies for Drug Development (Draft)." (2019).
[3] Food and Drug Administraion. "Framework for FDA's Real-World Evidence Program." (2018).
[4] Chen, Jie, et al. "The current landscape in biostatistics of real-world data and evidence: clinical study design and analysis." Statistics in Biopharmaceutical Research (2021): 1-14.
[5] Pocock, Stuart J. "The combination of randomized and historical controls in clinical trials." Journal of chronic diseases 29.3 (1976): 175-188.
[6] 武田健太朗ら. "臨床試験におけるヒストリカルコントロールデータの利用." 計量生物学 36.1 (2015): 25-50.

3.1.3　ヒストリカルコントロールを利用する解析手法

3.1.3.1　階層モデル

階層モデルによるアプローチでは，新規試験の対照群のパラメータとヒストリカルコントロールのパラメータを同じ分布からのランダムサンプリングと仮定する。これにより，新規試験の対照群のパラメータとヒストリカルコントロールのパラメータに，完全に同一でも完全に独立でもないという仮定を置くことができる[1]。ヒストリカルコントロールデータの借用度は試験間のばらつき（試験間分散）の大きさによる。試験間のばらつきがない場合はヒストリカルコントロールデータをすべて借用し，試験間のばらつきが大きくなるに従ってヒストリカルコントロールデータの借用度は小さくなる。

階層モデルによるアプローチは Meta-Analytic Predictive（MAP）と Meta-Analytic Combined（MAC）の 2 つに大別される。MAP は，ヒストリカルコントロールデータに基づいて予測分布を推定し，この予測分布を新規試験の対照群の事前分布として利用して，パラメータの推定を行う。新規試験データが得られていない段階でも事前分布の推定が可能であることから前向き（prospective）な方法と言える。一方で，MAC は，ヒストリカルコントロールデータに基づく事前分布を設定せず，得られた新規試験データとヒストリカルコントロールデータから直接事後分布を推定する方法である。基本的には，新規試験データが得られた後に適用することになるため，後ろ向き（retrospective）な方法と言える。このような背景から，前向きに試験デザインを考える際の階層モデルによるアプローチでは MAP を用いられることが多い。

本節では MAP アプローチの解析手法を以下で紹介する。なお，以下では，正規分布に従うデータを前提として記載するが，他の分布に従うデータであっても，要約統計量を正規近似することで同様に適用することができる[1]。

H 個のヒストリカルコントロールデータに対して，ヒストリカルコントロールデータ$h(=1,\ldots,H)$の要約統計量，及びパラメータをそれぞれY_h，及びθ_hとする。また，新規試験の対照群の要約統計量，及びパラメータをそれぞれY^*，及びθ^*とする。このとき，サンプリングモデル，及び初期事前分布は以下のように置くことができる。

サンプリングモデル

$$Y_h|\theta_h,\sigma_h \sim N(\theta_h,\sigma_h^2)$$

初期事前分布

$$\theta_1,\ldots,\theta_H,\theta^* \sim N(\mu,\tau^2)$$

ここで，τ^2は試験間分散，σ_h^2は試験内分散であり，本節では試験内分散σ_h^2は既知と仮定する。ヒストリカルコントロールと新規試験のパラメータを関連付けた初期事前分布にはいくつかの方法があるが[2, 3]，ここでは，すべてのパラメータが共通の分布に従うという最も単純な初期事前分布を利用する。

パラメータμ，及びτ^2の不確実性を考慮し，以下のような超事前分布を仮定する。ここで，$IGa(a_\phi, b_\phi)$は，形状パラメータa_ϕ，尺度パラメータb_ϕの逆ガンマ分布である。

超事前分布

$$\mu \sim N(\mu_\phi, \tau_\phi^2), \qquad \tau^2 \sim IGa(a_\phi, b_\phi)$$

新規試験の事前分布（MAP prior）として，ヒストリカルコントロールデータの予測分布$p(\theta^*|Y_1, \ldots, Y_H)$を利用することができる[4, 5]。予測分布は以下の式で表される。

$$p(\theta^*|Y_1, \ldots, Y_H) = \int p(\theta^*|\mu, \tau) \prod_h L(\theta_h|Y_h) p(\theta_h|\mu, \tau) p(\mu, \tau)\, d\theta_1 \cdots d\theta_H d\mu d\tau$$

ここで，τ^2，及びσ_h^2が既知の場合，μの超事前分布に一様分布を仮定すると，θ^*の予測分布は以下のようになる。

$$\theta^*|Y_1, \ldots, Y_H, \tau \sim N\left(\frac{\sum w_h Y_h}{\sum w_h}, \frac{1}{\sum w_h} + \tau^2\right), \qquad w_h = (\sigma_h^2 + \tau^2)^{-1}$$

前述の MAP prior による方法では，新規試験の対照群データとヒストリカルコントロールデータのパラメータの類似性とこれらのデータのパラメータの交換可能性（新規試験の対照群データとヒストリカルコントロールデータのパラメータ$\theta_1, \ldots, \theta_H, \theta^*$の順序を入れ替えても，同じ事前分布$p(\theta^*|Y_1, \ldots, Y_H)$が得られるということ）を仮定している。しかし，現実には，新規試験の対照群データとヒストリカルコントロールデータのパラメータが類似しておらず，これらのパラメータの交換可能性が成り立たない可能性も考えられる。このような場合を考慮した頑健な方法として，Robust MAP prior が提案されている[6]。MAP prior を$p_{MAP}(\theta^*)$，曖昧な事前分布（Vague prior）を$p_{vague}(\theta^*)$とすると，Robust MAP prior は以下の式で表される。

$$p_{R-MAP}(\theta^*) = (1 - w) p_{MAP}(\theta^*) + w p_{vague}(\theta^*)$$

ここで，wは新規試験の対照群データとヒストリカルコントロールデータのパラメータが類似していない事前確率と解釈することができ，ヒストリカルコントロールデータの借用度に対する懐疑性の程度を表している。Robust MAP prior は，MAP prior よりも裾を引いた分布である。そのため，新規試験の対照群とヒストリカルコントロールのデータのパラメータが類似していない場合，Robust MAP prior を用いることで，MAP prior を用いるときと比較して新規試験の尤度に近い事後分布が得られる。

また，新規試験の対照群データと複数のヒストリカルコントロールデータのパラメータの類似度に応じてヒストリカルコントロールデータの借用度を適応的に調整する方法として，Bayesian semiparametric MAP（Base-MAP）が提案されている[7]。Base-MAP では，MAP prior での試験間変動をディリクレ過程混合モデル（Dirichlet process mixture model）でノンパラメトリックにモデル化することで，新規試験の対照群データとヒストリカルコントロールデータのパラメータの類似性だけでなく，ヒストリカルコントロールデータ間のパラメータの類似性も考慮して，ヒストリカルコントロールデータの借用度を調整する。また，Base-MAP はヒストリカルコントロールデータの借用度の事前指定が不要であり，Robust MAP よりも柔軟な頑健法であるといえる。Hupf et al. (2021) [7]では，ヒストリカルコントロールデータ間の異質性や新規試験の対照群データとヒストリカルコントロールデータのパラメータの類似度を変化させたいくつかのシナリオでシミュレーションを実施して，Base-MAP と MAP や Robust MAP を含めたいくつかの解析手法の性能を比較しており，検討されたすべてのシナリオにおいて，第 1 種の過誤確率及び検出力の観点で，Base-MAP は MAP 及び Robust-MAP よりも良い性能を示した。

参考文献

[1] 武田健太朗ら. "臨床試験におけるヒストリカルコントロールデータの利用." 計量生物学 36.1 (2015): 25-50.

[2] Pocock, Stuart J. "The combination of randomized and historical controls in clinical trials." Journal of chronic diseases 29.3 (1976): 175-188.

[3] Turner, Rebecca M., et al. "Bias modelling in evidence synthesis." Journal of the Royal Statistical Society: Series A (Statistics in Society) 172.1 (2009): 21-47.

[4] Neuenschwander, Beat, et al. "Summarizing historical information on controls in clinical trials." Clinical Trials 7.1 (2010): 5-18.

[5] Gsteiger, Sandro, et al. "Using historical control information for the design and analysis of clinical trials with overdispersed count data." Statistics in Medicine 32.21 (2013): 3609-3622.

[6] Schmidli, Heinz, et al. "Robust meta-analytic-predictive priors in clinical trials with historical control information." Biometrics 70.4 (2014): 1023-1032.

[7] Hupf, Bradley, et al. "Bayesian semiparametric meta‐analytic‐predictive prior for historical control borrowing in clinical trials." Statistics in Medicine 40.14 (2021): 3385-3399.

3.1.3.2 Power prior

Power prior は初期事前分布と情報量を割り引かれたヒストリカルコントロールデータの尤度から得られた事後分布を指し，新規試験のデータを解析する際にヒストリカルコントロールデータの情報を有した事前分布として用いられる。関心のあるパラメータの初期事前分布と 1 つの過去の試験からのヒストリカルコントロールデータの尤度をべき乗パラメータ（$a_0 \in [0,1]$）でべき乗した項の積として定義される。a_0はヒストリカルデータから得られる情報量を調整しているべき乗パラメータである。a_0=0 の場合はヒストリカルコントロールデータを活用せず（No historical borrowing），a_0=1 の場合はヒストリカルコントロールデータをすべて活用する（Full historical borrowing）。したがって，ヒストリカルコントロールデータの信頼性，及びばらつきを考慮しながら，ヒストリカルコントロールデータの情報を事前分布に組み入れることができる。この点が，希少疾病の領域でも柔軟な薬効評価を可能にするメリットと考えられる。Power prior の算出方法として，これまでに多くの方法が提案されており，各方法の概要を以下で紹介する[1]。

i. Conditional power prior

Conditional power prior [2]はa_0に分布を仮定せずに既知の定数を与える方法である。$p_0(\theta)$を関心のあるパラメータθの初期事前分布，D_0をヒストリカルコントロールデータ，$L(\theta|D_0)$をヒストリカルコントロールデータの尤度としたとき，Conditional power prior は以下の通りである：

$$p(\theta|D_0, a_0) \propto L(\theta|D_0)^{a_0} p_0(\theta)$$

ii. Joint power prior

Joint power prior [2]はa_0を事前に特定することが難しい場合等にa_0の不確実性を考慮し，a_0に事前分布を仮定した上でヒストリカルコントロールデータの情報を調整する方法である。$p(a_0)$をa_0が従う事前分布（例えば，ベータ分布が考えられる）としたとき，Joint power prior は以下の通りである：

$$p(\theta, a_0|D_0) \propto L(\theta|D_0)^{a_0} p_0(\theta) p(a_0)$$

iii. Modified power prior

Modified power prior [3, 4]は Joint power prior の尤度原理（統計モデルのモデルパラメータに関連するデータのすべての情報が尤度関数に含まれる）を満たしていない点を補うために，基準化された Conditional power prior とべき乗パラメータa_0の事前分布の積で構成される。$\int L(\theta|D_0)^{a_0} p_0(\theta) d\theta$が基準化定数を表すとき，Modified power prior は以下の通りである：

$$p(\theta, a_0|D_0) \propto \frac{L(\theta|D_0)^{a_0} p_0(\theta)}{\int L(\theta|D_0)^{a_0} p_0(\theta) d\theta}\ p(a_0)$$

iv. Commensurate power prior

Commensurate power prior [5]はヒストリカルコントロールデータと新規試験の対照群データの類似性をパラメータ化し，ヒストリカルコントロールデータD_0と新規試験のデータDの共通性（Commensurability）に応じて，ヒストリカルコントロールデータの情報を調整する方法である。θと異なるθ_0をヒストリカルコントロールデータD_0から得られる関心があるパラメータとし，θの事前分布が平均θ_0，精度τ（分散の逆数：共通性の尺度）の正規分布に従うとする。θ_0の初期事前分布を$p_0(\theta_0)$，α_0の事前分布をτの情報を取り入れた$p(a_0|\tau)$，τの事前分布を$p(\tau)$としたとき，Commensurate power prior は以下の通りである：

$$p(\theta, a_0, \tau|D_0) \propto \frac{L(\theta_0|D_0)^{a_0} p_0(\theta_0)}{\int L(\theta_0|D_0)^{a_0} p_0(\theta_0) d\theta_0} p_0(\theta|\theta_0, \tau)\ p(a_0|\tau) p(\tau)$$

v. Multiple joint power prior

Multiple joint power prior [2]は複数の過去の試験データを新規試験に活用するため，Joint power prior を複数のヒストリカルコントロールデータに一般化し，H個のヒストリカルコントロールデータのそれぞれに対して，べき乗パラメータ$a_0 = (a_{01}, \cdots, a_{0H})$を互いに独立な事前分布に従うとした方法である。$h$番目のヒストリカルコントロールデータを$D_{0h}$，$a_{0h}$を$D_{0h}$のべき乗パラメータとし，互いに独立な事前分布に従うとしたとき，Multiple joint power prior は以下の通りである：

$$p(\theta, a_0|D_{01}, \ldots, D_{0H}) \propto \left\{\prod_{h=1}^{H} L(\theta|D_{0h})^{a_{0h}}\, p(a_{0h})\right\} p_0(\theta)$$

vi. Multiple modified power prior

Multiple modified power prior [3, 4]は上述の Modified power prior と同様に Multiple joint power prior の尤度原理を満たしていない点を補い，Modified power prior を複数のヒストリカルコントロールデータも扱えるように拡張した方法である。$C(a_1, \ldots, a_H)$を基準化定数を表すとしたとき，Multiple modified power prior は以下の通りである：

$$p(\theta, a_0|D_{01}, \ldots, D_{0H}) \propto \frac{1}{C(a_1, \ldots, a_H)} \left\{\prod_{h=1}^{H} L(\theta|D_{0h})^{a_{0h}}\, p(a_{0h})\right\} p_0(\theta)$$

vii. Dependent Modified Power Prior

Multiple joint power prior や Multiple modified power prior は複数のヒストリカルコントロールデータ間の類似性を考慮に入れていないが，利用可能な試験が Pocock の基準を満たす場合などでは，複数のヒストリカルコントロール間で関連するべき乗パラメータを求めることが適切な場面も考えられる。Dependent Modified Power Prior [6]は，複数のヒストリカルコントロールデータと新規試験の対照群のデータが完全に独立でないことを仮定し，べき乗パラメータa_{0h}が階層ベイズの枠組みで同じ分布に従うとした方法である。a_{0h}が超パラメータのα，βをもつベータ分布に従うとき，このベータ分布は平均$\mu = \frac{\alpha}{\alpha+\beta}$，分散$\sigma^2 = \frac{\mu(1-\mu)}{\alpha+\beta+1}$と再パラメータ化ができる。$\mu$の事前分布を$p(\mu)$，$\sigma^2$の事前分布を$p(\sigma^2)$としたとき，Dependent Modified Power Prior は以下の通りである：

$$p(\theta, a_{01}, \ldots, a_{0H}, \mu, \sigma^2 | D_{01}, \ldots, D_{0H}) \propto \frac{1}{C(a_1, \ldots, a_H)} \left\{ \prod_{h=1}^{H} L(\theta | D_{0h})^{a_{0h}} p(a_{0h} | \mu, \sigma^2) \right\} p_0(\theta) p(\mu) p(\sigma^2)$$

さらに，ヒストリカルコントロールデータと新規試験の対照群のデータで不整合があるときに，ベキ乗パラメータの分布を通してヒストリカルデータから借りる情報を減らすことを目的に$\pi(a_{0h}|\mu, \sigma^2)$と無情報事前分布を意図した分布との混合事前分布である Robust dependent modified power prior も提案されている。

参考文献

[1] 武田健太朗ら. "臨床試験におけるヒストリカルコントロールデータの利用." 計量生物学 36.1 (2015): 25-50.

[2] Ibrahim, Joseph G., and Ming-Hui Chen. "Power prior distributions for regression models." Statistical Science (2000): 46-60.

[3] Duan, Yuyan, Keying Ye, and Eric P. Smith. "Evaluating water quality using power priors to incorporate historical information." Environmetrics: The Official Journal of the International Environmetrics Society 17.1 (2006): 95-106.

[4] Neuenschwander, Beat, Michael Branson, and David J. Spiegelhalter. "A note on the power prior." Statistics in medicine 28.28 (2009): 3562-3566.

[5] Hobbs, Brian P., et al. "Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials." Biometrics 67.3 (2011): 1047-1056.

[6] Banbeta, Akalu, et al. "Modified power prior with multiple historical trials for binary endpoints." Statistics in Medicine 38.7 (2019): 1147-1169.

3.1.3.3 マッチング（Matching）

外部試験には新規試験とは背景の異なる被験者が含まれる可能性があり，治療結果の推定量にバイアスが生じうる。そのため，バイアスのない推定量を求めるため，新規試験データと外部試験データの患者の背景因子（ベースライン共変量）の不均一性を調整する方法として傾向スコア等のマッチングの活用が考えられる。

新規試験データの患者について，調整したい背景因子に完全に一致するような外部試験データの患者を選ぶ方法を単純なマッチングと呼ぶ。単純なマッチングを用いた評価を行った事例は「3.1.5.1 **セルリポナーゼ アルファ（ブリニューラ®）**」で紹介する。一方，単純なマッチングで考慮できる因子数は限られるため，傾向スコアを使ったマッチングがよく用いられる。

傾向スコアは Rosenbaum & Rubin (1983) [1]によって提案された，複数の交絡因子を調整し，因果効果を推定するための方法である。ここでは，単群試験または比較試験の対照群に対して，外部試験データを活用して，新規試験データを補う目的で傾向スコアを用いる。$Z = 1$が新規試験データ，$Z = 0$が外部データを表すとする。$\boldsymbol{X}$をベースライン共変量とするとき，傾向スコアは以下で表される：

$$e(\boldsymbol{X}) = \Pr\left[Z = 1|\boldsymbol{X}\right]$$

傾向スコアを用いて因果効果を推定するための前提条件が「強く無視できる割り当て」[2]（岩崎学 (2015)）であり，このとき，傾向スコアの各値のもとでの新規試験データと外部試験データの治療結果はバイアスのない推定量になる。

以下では，マッチングの枠組みで，外部データを活用し，新規試験の治療結果を推定する方法の概要を紹介する。

i. 傾向スコアに基づく MAP prior アプローチ

傾向スコアをいくつかの層に分割し，層内の新規試験データと外部試験データの傾向スコアの類似性を評価することで，層間の不均一性を考慮した上で，層ごとに治療結果を推定し，最後に併合することで全体の治療結果を求める方法が提案されている。ここでは，MAP prior に基づく方法について説明する[3]。

Step 1:傾向スコアの推定，及び切り取り（Trimming）

各被験者 i の傾向スコア$\hat{e}(\boldsymbol{X_i})$を求める。$E_1 = \{\hat{e}(\boldsymbol{X_i}): Z_i = 1\}$を新規試験の患者から推定した傾向スコアの集合とする。新規試験の患者と類似する傾向スコアを持つ外部試験データの患者の部分集団を選択するため，E_1の範囲外の傾向スコアをもつ外部試験データの患者を除外する。

Step 2:層別化

Step 1 で得られた患者集団に対して，傾向スコアによる層別化を行い，S個の層に分ける。s で s 番目の層を示し（$s = 1, \ldots, S$），s番目の層に該当する傾向スコアの患者の部分集団を$X_s = \{\boldsymbol{X}: \hat{e}(\boldsymbol{X}) \in (\hat{q}_{s-1}, \hat{q}_s]\}$（ベースライン共変量），$D_{s,Z} = \{(y_i, \boldsymbol{X_i}): \boldsymbol{X_i} \in X_s, Z_i = z\}$で表すとする。また，

s番目の層に含まれる外部試験，及び新規試験のデータからの患者の数は$n_{s,0}$，$n_{s,1}$とする。

Step 3：層特有の MAP prior

s 番目の層の$Y_{s,Z} = \{y_i : \boldsymbol{X_i} \in X_s, Z_i = z\}$ を連続量の結果としたとき，ベイズ流階層モデルは以下の通りである。

$$Y_{s,z}|\theta_{s,z} \sim \mathrm{N}\left[\theta_{s,z}, s_{s,z}^2\right], \qquad s = 1, \dots, S, z = 0,1$$
$$\theta_{s,z}|\mu_s, \tau_s \sim \mathrm{N}[\mu_s, \tau_s^2], \qquad s = 1, \dots, S, z = 0,1$$

層特有の効果μ_sは曖昧な事前分布（Vague prior），層特有の標準偏差τ_sは半正規事前分布に従うとする。層特有の MAP prior として，各層で外部データと新規試験データは同じ分布に従うと仮定する。各層での不均一性を分散パラメータτ_s^2で調整する。このとき，外部試験データを活用した新規試験データの全体の治療結果θ_1は層特有の MAP prior の治療結果の重み付き平均となる。

$$\theta_1 = \sum_{s=1}^{S} \frac{n_{s,1}}{N_1} \theta_{s,1}, N_1 = \sum_{s=1}^{S} n_{s,1}$$

Step 4：傾向スコアに基づく MAP prior（PS-MAP prior）

パラメータτ_sは各層で外部試験データと新規試験データの間のベースライン共変量の類似性を反映している。一方で，各層の類似性に関する情報を事前に知ることは難しく，各層の分散に関する情報も共有することが望ましい。したがって，パラメータτ_sの事前分布を以下と定義する。

$$\tau_s \sim \text{half normal}[k_s \times t]$$

このとき，k_sは各層の類似性を反映している。また，tは層特有の分散をスケール化させる「ベースライン」分散を反映させ，事前の Effective sample size (Prior ESS)に基づき更新される[4, 5]。Prior ESS は事前分布によって与えられる情報量を定量化する方法であり，外部データから情報を借りる症例数を臨床家や試験チームメンバーなどと相談した上で決定する。

各層の外部試験データと新規試験データのベースライン共変量の類似性を測る指標として，外部試験データと新規試験データの患者のs番目の層の傾向スコアの密度関数$f_{s,0}$，及び$f_{s,1}$の曲線が重複している領域（Overlapping coefficients）を以下の通りに求める。

$$r_s = \int_0^1 \min\left[f_{s,0}(e), f_{s,1}(e)\right] de$$

S個の層で Overlapping coefficients, $R = (r_1, r_2, \ldots, r_S)$, を計算したあと, r_sに反比例する$k_s = r^{ref}/r_s$を求める。r^{ref}としてRの中央値等を用いる。次に$t^{(i)}$の初期値として, $t^{(1)} = 1$とする。このとき, τ_sの事前分布が定まり($\tau_s \sim \text{half normal}[k_s * t^{(i)}]$), PS-MAP prior が求まる。つまり, 傾向スコアに基づき, 外部試験データと新規試験データの類似性が高いと, k_sとともにτ_sが従う半正規分布の分散が小さくなり, 外部試験データからより多くの情報を借りることができる。PS-MAP prior の Prior ESS が事前に特定した範囲に収まる場合, PS-MAP prior が決定し, 収まらない場合, $t^{(i)}$を更新し, Prior ESS が事前に特定した範囲に収まるまで, PS-MAP prior の計算を繰り返す。

Step 5:解析結果

新規試験の結果が$p(Y_1|\theta_1)$によってモデル化され, PS-MAP prior が階層モデルと層特有の治療効果の重み付き平均から$p_{\text{PS-MAP}}(\theta_1) = \boldsymbol{p}(\theta_1|X_0, X_1, Y_0)$($X_0$は外部試験データの共変量, Y_0は外部試験データの目的変数)として算出したとき, 新規試験データの関心のあるパラメータθ_1の事後分布は以下で求められる。

$$p(\theta_1|Y_1) \propto p(Y_1|\theta_1)\, p_{\text{PS-MAP}}(\theta_1)$$

ii. 傾向スコアに基づく Power prior アプローチ

i では, 新規試験データと外部試験データの傾向スコアに基づく各層の不均一性を MAP prior の枠組みで調整する方法について紹介したが, ここでは Power prior の枠組みで調整する方法を紹介する[6]。Step 1, 及び Step 2 は PS-MAP prior の手順と同じである。

Step 3':層特有の Power prior

新規試験の治療結果を示すs番目の層の関心のあるパラメータを$\theta_{s,1}$の分布は, 層特有の Power prior α_s ($0 \le \alpha_s \le 1$)を用いて, 以下で表すことができる。

$$p(\theta_{s,1}) \propto L(\theta_{s,1}|D_{s,0})^{\alpha_s} p_0(\theta_{s,1})$$

Step 4':傾向スコアに基づく Power prior(PS-power prior)

α_sは Conditional power prior として以下で定義する。

$$\alpha_s = \min\left(\frac{A \times v_s}{n_{s,0}}, 1\right), \qquad v_s \ge 0, \sum_{s=1}^{S} v_s = 1, A \in [0, N_0]$$

Aは外部試験データから情報を借りる全体の目標症例数を示し, 解析前に事前に特定する。Av_sは第s層において情報を借りる目標症例数, v_sは第s層の外部試験データがθ_sに寄与する情報

量を意味する。v_sを特定する方法としては，以下の 2 種類が考えられる。

(1) 比例配分：AをS個の層に配分し，標準化した Overlapping coefficients を以下で求める。

$$v_s = \frac{r_s}{\sum_{s=1}^{S} r_s}$$

このとき，PS-power prior は以下となる。

$$p(\theta_{1,1}, \ldots, \theta_{S,1}) \propto \prod_{s=1}^{S} L(\theta_{s,1}|D_{s,0})^{\alpha_s} \, p_0(\theta_{1,1}, \ldots, \theta_{S,1})$$

(2) ベイズ流配分：$v_1, \ldots, v_S$が以下のディリクレ事前分布に従うとする。ここでRはv_sの事前分散を制御する定数である。

$$p_0(v_1, \ldots, v_S) = \mathrm{Dir}\left(\frac{r_1}{R}, \ldots, \frac{r_S}{R}\right)$$

このとき，PS-power prior は以下となる。

$$p(\theta_{1,1}, \ldots, \theta_{S,1}, v_1, \ldots, v_S) \propto \frac{\prod_{s=1}^{S} L(\theta_{s,1}|D_{s,0})^{\alpha_s}}{\int \prod_{s=1}^{S} L(\theta_{s,1}|D_{s,0})^{\alpha_s} \, d\theta_{1,1} \ldots d\theta_{S,1}} p_0(\theta_{1,1}, \ldots, \theta_{S,1}) p_0(v_1, \ldots, v_S)$$

Step 5’：解析結果

s番目の層の関心のあるパラメータを$\theta_{s,1}$の事後分布は以下で求まる。

(1) 比例配分：

$$p(\theta_{1,1}, \ldots, \theta_{S,1}|D_1) \propto \prod_{s=1}^{S} L(\theta_{s,1}|D_{s,1}) \, p(\theta_{1,1}, \ldots, \theta_{S,1})$$

(2) ベイズ流配分：

$$p(\theta_{1,1}, \ldots, \theta_{S,1}, v_1, \ldots, v_S|D_1) \propto \prod_{s=1}^{S} L(\theta_{s,1}|D_{s,1}) \, p(\theta_{1,1}, \ldots, \theta_{S,1}, v_1, \ldots, v_S)$$

最後にs番目の層の関心のある治療結果$\theta_{s,1}$を併合し，θ_1の事後分布を以下から算出する。

$$\theta_1 = \sum_{s=1}^{S} (\dot{q}_s - \dot{q}_{s-1})\, \theta_{s,1}$$

$\dot{q}_s$は重み付き平均等（$\dot{q}_s = s/S$）で定義する。

参考文献

[1] Rosenbaum, Paul R., and Donald B. Rubin. "The central role of the propensity score in observational studies for causal effects." Biometrika 70.1 (1983): 41-55.

[2] 岩崎学. 統計的因果推論. 朝倉書店, 2015.

[3] Liu, Meizi, et al. "Propensity‐score‐based meta‐analytic predictive prior for incorporating real‐world and historical data." Statistics in medicine 40.22 (2021): 4794-4808.

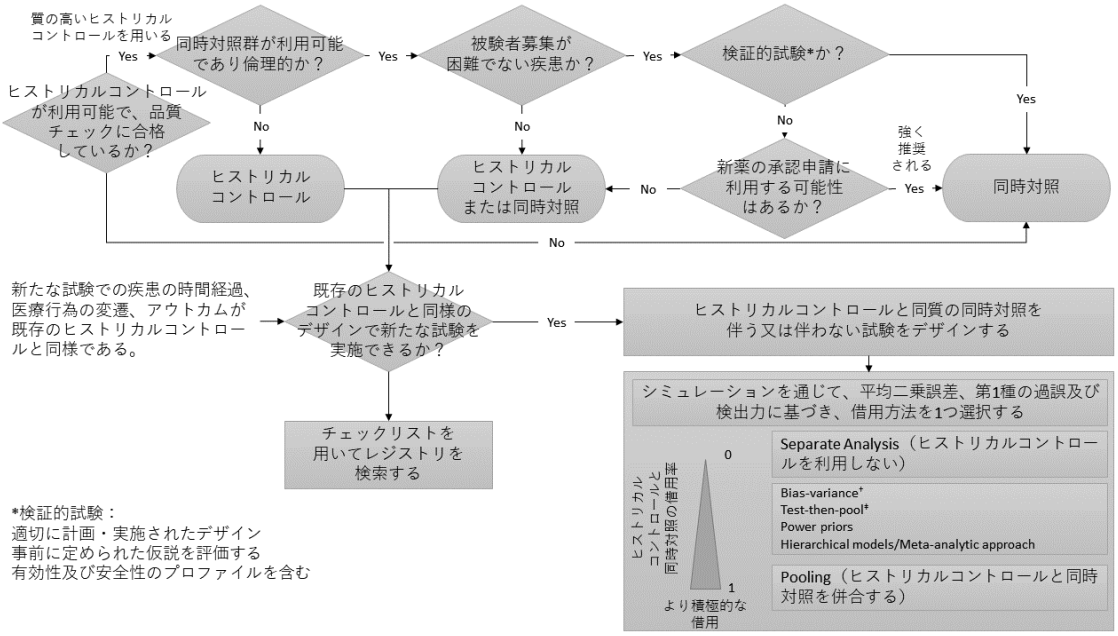
[4] Morita, Satoshi, Peter F. Thall, and Peter Müller. "Determining the effective sample size of a parametric prior." Biometrics 64.2 (2008): 595-602.

[5] Neuenschwander, Beat, et al. "Predictively consistent prior effective sample sizes." Biometrics 76.2 (2020): 578-587.

[6] Wang, Chenguang, et al. "Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies." Journal of Biopharmaceutical Statistics 29.5 (2019): 731-748.

3.1.4　ヒストリカルコントロールの解析手法適用時の留意点

臨床試験でヒストリカルコントロールデータを利用するにあたって，同時対照群の設定や解析手法を検討する必要がある。Ghadessi et al. (2020) [1]は，臨床試験でヒストリカルコントロールを利用するための Decision making diagram を提案しており，これらの検討に役立つ。



†Bias-variance は新規試験での対照群データとヒストリカルコントロールデータの違いをバイアスパラメータとして扱うことで，これらのデータ間の差異を考慮しながらベイズ推定する方法
‡Test-then-pool は新規試験での対照群データとヒストリカルコントロールデータの類似性を仮説検定により判定し，帰無仮説が棄却されなかった場合は併合し，棄却された場合は新規試験データのみを用いる方法

図 3-1-4-1 臨床試験でヒストリカルコントロールを利用するための Decision Making Diagram（Ghadessi et al. (2020) [1] 改編）

解析手法に関するシミュレーションを実施し，解析手法の特性（弱み，強み）を把握しておくことは解析手法の選択の際に有益である。シミュレーションでは，あるデザインでの最適なモデルを選択するための様々な方法の比較ができる。すなわち，ヒストリカルコントロールデータを借用する場合に，「スイートスポット」と呼ばれる解析手法が最も威力を発揮する（平均二乗誤差が小さく，第 1 種の過誤確率が低く，検出力が高い）範囲を比較する。同時対照群とヒストリカルコントロールのデータが類似している場合は，ヒストリカルコントロールデータを借用することで，第 1 種の過誤確率が減少し，検出力が向上する。反対に，同時対照群とヒストリカルコントロールのデータが類似していない場合は，検出力が低下することや，第 1 種の過誤確率が増大することがある。既知の外部情報や潜在的バイアスのために第 1 種の過誤確率が増大するかもしれないことに規制当局としては懸念があるだろう[2]。そのため，ヒストリカルコントロールを利用する際の解析手法の動作特性の評価は，設定が適切であるかを判断する際の重要な材料となる。

ヒストリカルコントロールを利用する際の重要なステップの一つとして，感度分析を実施し，主解

析の結果の安定性（ロバストネス）を評価することが挙げられる。感度分析の際に考慮する事項として以下が考えられる[1]。

- 選択バイアスの影響
- 主要な結果に対する欠測データの影響
- ヒストリカルコントロールと新規試験データのマッチング方法
- 異なる事前分布や方法を用いた場合のヒストリカルコントロールの借用度
- 重み付け係数の変更
- 複数のヒストリカルコントロールを用いる場合に，一部のヒストリカルコントロールを除いたときの影響
- Pocock の基準を満たしていないヒストリカルコントロールの影響
- （ヒストリカルデータが複数ある場合の）ヒストリカルコントロールデータ間の異質性，及びヒストリカルコントロールデータと同時対照データの異質性

上記の点を踏まえて，ヒストリカルコントロールを利用する際に，3.1.3 節のどの解析手法を用いるべきかについては，得られている情報や新規試験の目的によって異なる。ヒストリカルコントロールとして活用できる過去試験が少数の場合には，階層モデルでは試験間分散の正しい推定が困難となり，Power prior の方がより適切な治療効果を得られるかもしれない。一方，過去試験が複数存在し，試験間に相関があると考えられる場合には，階層モデルの方が自然な枠組みの下で適切な治療効果が得られるかもしれない。また，ヒストリカルコントロールの患者の背景因子の情報が利用可能な場面では，患者の背景因子の不均衡を傾向スコアで調整することでバイアスのない治療効果が推定できると考えられる。このようにヒストリカルコントロールを活用するための計画においては，（1）解析手法の選択前にヒストリカルコントロールが利用可能かの判断，（2）新規試験開始前に選択する解析手法についての動作特性の把握，（3）新規試験データが得られた後の解析手法の適切性を評価する感度解析，を行うことが重要である。

参考文献

[1] Ghadessi, Mercedeh, et al. "A roadmap to using historical controls in clinical trials–by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG)." Orphanet Journal of Rare Diseases 15.1 (2020): 1-19.

[2] Chen, Jie, et al. "The current landscape in biostatistics of real-world data and evidence: clinical study design and analysis." Statistics in Biopharmaceutical Research (2021): 1-14.

3.1.5　ヒストリカルコントロールを利用した事例

本節では，ヒストリカルコントロール群を用いた臨床評価に関して，4 品目の承認審査事例を取り上げる。有効性評価の妥当性が議論された 3 品目と試験デザインを工夫することで薬効評価の比較可能性を担保した 1 品目である。取り上げた事例は，いずれも希少疾患であり，かつ，結果として小児を対象として開発された薬剤であった。これらの事例では，プラセボを対照群としたランダム化比較試験を行うことは困難な状況であった。また，治験依頼者がヒストリカルコントロール群となるデータを集積することも困難であったと思われ，規制当局にとっても難しい審査であったと思われる。本節で取り上げる事例についての簡潔な要約を以下の表 3-1-5-1 に示す。

なお，薬事申請における Real World Data の利活用については，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の「薬事申請に Real World Data を外部対照として利用する際の留意点」[1]で詳細に議論されているので参考にされたい。傾向スコアを用いた事例も，当該文書でいくつか報告されている。

表 3-1-5-1 本節で取り上げるヒストリカルコントロールを用いた事例

薬剤名 （該当する節）	対象疾患	規制当局	事例としての特徴
セルリポナーゼアルファ （3.1.5.1 節）	セロイドリポフスチン症 2 型	FDA， PMDA	自然歴研究をヒストリカルコントロール群として設定したものの，評価指標，及び患者背景特性の違いについて指摘された事例。主要評価項目の変更や追加解析を規制当局から求められたが，最終的にはヒストリカルコントロールとの比較の結果に基づき承認された。
nifurtimox （3.1.5.2 節）	アメリカトリパノソーマ症（シャーガス病）	FDA	ヒストリカルコントロール群が設定されたが，主要評価項目の臨床的意義について審査過程で懸念が示され，ヒストリカルコントロールとの比較に基づく薬効評価が困難となった事例。
eteplirsen （3.1.5.3 節）	デュシェンヌ型筋ジストロフィー	FDA	長期的な薬効に関して，継続長期投与試験とヒストリカルコントロール群のデータを比較して評価することを試みたものの，規制当局からは否定的な見解が得られた事例。
ビルトラルセン （3.1.5.4 節）	デュシェンヌ型筋ジストロフィー	PMDA	試験で収集する評価項目をレジストリの収集項目に合わせることによって薬効評価の比較可能性を高めた事例。

参考文献

[1] 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. "薬事申請に Real World Data を外部対照として利用する際の留意点."(2022).

3.1.5.1　セルリポナーゼ アルファ（ブリニューラ®）

セロイドリポフスチン症 2 型（CLN2）について

CLN2 は，ライソゾーム内のセリンプロテアーゼであるトリペプチジルペプチダーゼ 1（TPP1）の欠損を特徴とする，極めてまれな進行性の神経変性を伴う重度の遺伝子疾患であり，日本では難病指定されている。CLN2 遺伝子の変異によって TPP1 が欠損すると，ライソゾーム内で代謝されるべき老廃物が多くの器官で細胞内に蓄積し，中枢神経で蓄積した場合は神経変性症状が引き起こされる。CLN2 は通常，2~4 歳の間に痙攣発作や運動失調，並びに言語発達遅延を伴って発症し，年齢を重ねるとともに多様な徴候が現れ，悪化の一途をたどる。末期では，失明，寝たきり，及びコミュニケーション不能となり，多くの場合は若年で死亡に至る。最初の症状発現時から死亡までの期間の中央値は 7.8 年という報告がある。欧米における CLN2 の推定発症率は出生児 10 万人当たり 0.5~1 人，推定有病率は 100 万人当たり 0.6~0.7 人であると報告されている。2019 年時点で確定診断されている日本人の CLN2 患者は 3 例のみであった。CLN2 の治療法は長らく疾患進行の安定化を図るための対症療法，及び緩和療法のみであった。

セルリポナーゼ　アルファについて

セルリポナーゼ　アルファは，チャイニーズハムスター卵巣細胞から産生される遺伝子組み換えヒト TPP1 であり，TPP1 欠損症として知られている CLN2 治療を目的とした酵素補充療法製剤である。通常，セルリポナーゼ　アルファとして，300 mg を 2 週間に 1 回，脳室内投与する。

セルリポナーゼ　アルファは「CLN2 の治療」の効能・効果で 2013 年 3 月に欧州で，2013 年 4 月に米国で，2018 年 9 月に日本で，希少疾病用医薬品に指定された。2017 年 4 月に米国で初めて承認され，2017 年 5 月に欧州で，その後，ウクライナ，ブラジル，及びオーストラリアで承認され，日本では 2019 年 9 月に承認された。

セルリポナーゼ　アルファの臨床試験

海外でセルリポナーゼ　アルファの有効性評価に用いられた臨床試験は，第 I/II 相試験の 190-201 試験，及び 190-202 試験のみであり，190-201 試験は実薬単群の非盲検用量漸増期（4～22 週間）と固定用量投与期（48 週間）から構成され，190-202 試験はその延長投与試験（最長 239 週）である。両試験は米国，ドイツ，イタリア，イギリスで実施され，24 例の CLN2 患者が参加し，日本人患者はドイツにて 1 例参加した。

開発会社の BioMarin 社が FDA に申請した有効性の主要評価項目は，投与 48 週後に，CLN2 評価尺度の運動・言語尺度がベースラインから 2 点以上低下しない，またはベースラインが 1 点の場合は 0 点にならない割合であった。CLN2 評価尺度は，運動，及び言語それぞれに 0~3 点のスコアを割り当て，その合計スコアを算出し，最終的に 0 点（重い障害）～6 点（正常）のスコアとした。脳室内投与用デバイスの植込み手術の実施や倫理的な観点から，同時対照群を設定することが困難であったため，190-901 試験として，DEM-CHILD データベースに登録されている CLN2 患者

69 例を CLN2 評価尺度で評価観察した自然歴研究を行い，190-201/202 試験の結果と比較した。190-201/202 試験の症例数算出に当たり，ドイツ・ハンブルグの自然歴データベースから遺伝子型が証明された CLN2 患者 30 人の長期的な解析から，患者は発症から 18 か月間で CLN2 評価尺度が平均 2 スコア，標準偏差 1.8 で低下すると推測された。治療による 18 か月間の平均変化が 0.5 スコア，ばらつきが同程度であれば，検出力 90%，両側有意水準 0.05 と仮定した場合，必要な症例数は 18 例，これに中止率 20%を加味して，22 例と設定した[1]。

自然歴研究との CLN2 評価尺度の違いに関する FDA からの指摘

FDA から，CLN2 評価尺度に関する 190-201/202 試験と 190-901 試験の 3 つの違いについて指摘された[2]。

① 評価尺度の違いについて

190-201/202 試験では，オリジナルの CLN2 評価尺度[3]から評価をより明確化して改良した尺度を使用したが，190-901 試験では，オリジナルの尺度を使用した。この違いについて表 3-1-5-1-1 に示す。FDA から，両者は異なるのではないか，特に言語尺度で 190-201/202 試験の方のスコアが高くなり，試験治療に有利にはたらくのではとの指摘があった。

表 3-1-5-1-1：190-901 試験（自然歴研究），及び 190-201/202 試験で用いた CLN2 評価尺度

	スコア	190-901 試験	190-201/202 試験
運動	3	正常に歩行	正常に歩行 すなわち際立った運動失調または病的な転倒がない
	2	頻回の転倒，明らかな巧緻運動障害	独歩は可能（介助なしに 10 歩以上歩ける）であるが，明らかに不安定な歩き方，頻回に転倒することがある
	1	介助なしには歩行不能または四つ這いのみ	介助なしには歩行できないまたは四つ這いのみ
	0	運動不能，ほぼ寝たきり	歩行または四つ這いができない
言語	3	正常	正常 すなわち明瞭かつ年齢相応の言語を話し，言語能力の低下はまだ認められない
	2	明らかな異常を認める	明らかな異常を認める すなわち理解しにくい単語がある ただし，考え，要望，必要なものを短い文で伝えることはできる 本スコアは，これまでの獲得言語（それぞれの患者

	スコア	190-901 試験	190-201/202 試験
			が最も獲得した到達点)からの低下を意味する
	1	ほとんど理解不能	ほとんど理解不能 すなわち理解できる言葉をほとんど発しない
	0	理解不能または発語なし	理解できる言葉及び発声がない

ブリニューラ脳室内注射液 150mg に関する資料[4] 表 2.7.3.1-2 をもとに作成

② 評価方法について

190-201/202 試験では撮影されたビデオを，予め定められた 4 人の医師のうち 1 人の医師が評価したが，同じ患者を常に同じ医師が評価するとは限らなかった。それに対して 190-901 試験はほとんどが保護者へのインタビューまたはカルテレビューによる後ろ向きな評価であったものの，前向きと後ろ向きな方法で収集されたデータが混在しており，同じ被験者でも時点によって収集される方法が異なる場合があった。

③ 評価のタイミングについて

190-201/202 試験では評価時点を試験で規定していたが，190-901 試験では不定期にならざるを得なかった。

これら 3 つの違いのため，FDA から CLN2 評価の再評価が求められ，190-201/202 試験で撮影されたビデオを，190-201/202 試験の trainer，及び 190-901 試験の CLN2 developer で再評価し，190-201/202 試験で得られた結果と比較した。その結果を表 3-1-5-1-2 に示す。運動尺度の評価はほぼ一致したのに対し，言語尺度の評価は一致しない傾向がみられたことから，FDA は運動尺度のみで試験の評価を実施するよう推奨した。

表 3-1-5-1-2：190-201/202 試験における CLN2 評価ビデオを，190-201/202 trainer または 190-901 CLN2 developer が評価した結果との比較

比較	運動尺度 重み付きカッパ係数	言語尺度 重み付きカッパ係数
190-201/202 試験結果 vs 190-201/202 trainer による評価	0.93	0.82
190-201/202 試験結果 vs 190-901 CLN2 developer による評価	0.88	0.53
190-901 CLN2 developer による評価 vs 190-201/202 trainer による評価	0.94	0.56

FDA STATISTICAL REVIEW [2] Table 5 をもとに作成

自然歴研究との患者背景の違いに関する FDA からの指摘

190-901 試験では，DEM-CHILD データベースに登録されている 69 例から，以下の基準を全て満たす患者 42 例を，190-201/202 試験と比較可能な患者として選定した。

・CLN2 評価スコアが利用できる患者

・190-201/202 試験に参加しなかった患者

・一卵性双生児を除いた患者

・最初の CLN2 評価から 6 か月以上経過した後に，1 回以上評価された患者

ただし，FDA からは 190-201/202 試験と 190-901 試験のベースラインの患者背景について，①男女比，②遺伝子変異を有する患者の割合，③出生年代の 3 つの違いから有効性評価に影響を与えるのではとの指摘があった。

表 3-1-5-1-3：190-901 試験（自然歴研究），及び 190-201/202 試験の患者背景のちがい

	190-901 試験 (n=42)	190-201/202 試験 (n=22)
性別		
男性	25 (60%)	7 (32%)
女性	17 (40%)	15 (68%)
疾患対立遺伝子		
一般的な対立遺伝子の 2 つがある	24 (57%)	9 (41%)
一般的な対立遺伝子の いずれか一方のみがある	11 (26%)	6 (27%)
一般的な対立遺伝子がない	7 (17%)	7 (32%)
出生年代		
1980 年より前	4 (10%)	0
1980 年代	2 (5%)	0
1990 年代	19 (45%)	0
2000 年代	16 (38%)	12 (55%)
2010 年以降	1 (2%)	10 (45%)

FDA STATISTICAL REVIEW [2] Table 16 をもとに作成

そこで，マッチング解析が実施された。マッチングは 1:1 で行われ，マッチングの基準は，ベースラインの運動尺度スコア，年齢（±3 か月），及び遺伝子変異型を用いて実施された。1 対複数または複数対 1 で一致した場合は，マッチングに用いる変数を，詳細な遺伝子変異型，性別，最初の症状が現れた年齢の順に追加した。その結果，17 組の患者が選ばれた。なお，スポンサー側は年

齢のマッチングに当初 12 か月の許容幅を提案したが，FDA から幅が広すぎることから 3 か月の許容幅を使うことを推奨されている。

FDA が推奨した主要評価項目は，マッチング解析で選ばれた 17 例で，CLN2 評価尺度の運動尺度のみがベースラインから 2 点以上低下しない，またはベースラインが 1 点の場合は 0 点にならない被験者の割合を，投与後 48 週よりも投与継続した後の 96 週で評価することであった。結果として 190-201/202 試験で，190-901 試験と比較して，CLN2 運動尺度の悪化抑制が認められた。

なお，添付文書には，言語尺度に関しては自然歴研究との比較ができなかったため，言語尺度の有効性は確立していないと記載されている[5]。

表 3-1-5-1-4：CLN2 評価尺度の運動尺度がベースラインから 2 点以上低下しないまたはベースラインが 1 点の場合は 0 点にならない症例数とその割合

投与期間	190-901 試験 (n=17) n (%)	190-201/202 試験 (n=17) n (%)	差 % (95%CI)	オッズ比 OR (95%CI)
48 週	13 (76%)	16 (94%)	18% (-19%, 51%)	0.25 (0.005, 2.53)
72 週	11 (65%)	16 (94%)	29% (-7%, 61%)	0.17 (0.004, 1.37)
96 週	6 (35%)	16 (94%)	59% (24%, 83%)	0.09 (0.002, 0.63)

FDA STATISTICAL REVIEW [2] Table 17 をもとに作成

日本におけるセルリポナーゼ　アルファの承認

日本の承認申請時には，190-901 試験から新規に患者を追加した集団で 190-901 試験補遺解析として後ろ向きの自然歴研究を行った。190-201/202 試験と比較するにあたり，1:1 マッチングにより，年齢について両試験の月齢差が 12 か月以下で，かつベースラインの運動・言語尺度が一致している 21 組の被験者を選択し，投与後 48 週における CLN2 評価尺度の運動尺度と言語尺度の合計スコアで比較を行った。合計スコアの低下が 48 週間あたり 2 点未満である患者をレスポンダーと定義した場合，その推定割合は 190-201/202 試験では 100%（21/21 例），190-901 試験補遺解析では 48%（10/21 例）であり，両試験間の推定値の差は 52%であった（p=0.0002）。機構からは，この自然歴研究との比較評価については限界があるものの，セルリポナーゼ　アルファにより疾患の進行が抑制される傾向が示唆されているとコメントされ，承認に至っている[6]。

まとめ

本事例は，自然歴研究を外部対照群として設定したものの，評価指標，及び患者背景の違いに

ついて指摘され，主要評価項目の変更や追加解析を求められた事例として紹介した。本事例の自然歴研究は後ろ向きの評価であったため，臨床試験と単純に比較することが困難であったと考えられる。治療法の確立していない希少疾病用医薬品開発において，疾患の自然経過に対して有意な治療効果を示すことは重要であるが，評価指標や患者背景に関する比較可能性を可能な限り高める試験デザインや解析計画が要求される。

参考文献

[1] Schulz, Angela, et al. "Study of intraventricular cerliponase alfa for CLN2 disease." New England Journal of Medicine 378.20 (2018): 1898-1907.

[2] Food and Drug Administration. “Brineura (cerliponase alpha) Injection Statistical Review and Evaluation.” (2017). https://www.accessdata.fda.gov/drugsatfda_docs/nda/2017/761052Orig1s000StatR.pdf

[3] Steinfeld, Robert, et al. "Late infantile neuronal ceroid lipofuscinosis: quantitative description of the clinical course in patients with CLN2 mutations." American journal of medical genetics 112.4 (2002): 347-354.

[4] BioMarin Pharmceutical Japan 株式会社. “ブリニューラ脳室内注射液 150mg に関する資料.” 医薬品医療機器総合機構（2019） https://www.pmda.go.jp/drugs/2019/P20191010003/index.html

[5] Food and Drug Administration. “HIGHLIGHTS OF PRESCRIBING INFORMATION.” (2017). https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/761052lbl.pdf

[6] 医薬品医療機器総合機構. “ブリニューラ脳室内注射液 150mg，審査報告書.” (2019). https://www.pmda.go.jp/drugs/2019/P20191010003/641173000_30100AMX00236_A100_2.pdf

3.1.5.2　nifurtimox（LAMPIT®）

アメリカトリパノソーマ症（シャーガス病）について[1]

アメリカトリパノソーマ症（シャーガス病）は昆虫媒介疾患（感染病）である。2006 年時点でラテンアメリカにおいて約 800 万人の感染者がおり，年間の新規患者数は 55,185 名と算出されている。シャーガス病はラテンアメリカ，及びカリブ地域に固有であると長い間考えられてきたが，世界的疾患になりつつあることを示唆するエビデンスが増えてきている。寄生虫曝露後の疾患の自然経過は，急性期では無症候性の可能性が高い。症状が現れた場合，通常は自然に軽快する発熱性の病気で感染者の約 90%で自然に消散するが，5%~10%は急性期に重度の心筋炎，及び/または髄膜脳炎により死亡する。新たな感染症のほとんどは 15 歳未満の小児で発生し，媒介生物または垂直感染によって起こり，1~5 歳の幼児で最も頻度が高い。

nifurtimox を含むシャーガス病の治療薬について[1]

2018 年時点で，南米の一部の国では，小児，及び成人におけるシャーガス病の治療薬として nifurtimox，及び benznidazole のみが承認されており，benznidazole は米国でも 2~12 歳の子供のシャーガス病の治療薬として承認されている。1960 年代後半に開発されたニトロフラン化合物である nifurtimox は，還元酸素代謝産物（例:スーパーオキシド，及び過酸化水素）の産生を介して寄生虫に作用すると考えられ，1970 年代前半に開発されたニトロイミダゾール誘導体である benznidazole は，ニトロ還元中間体の寄生虫分子への共有結合を介して作用すると考えられている。

nifurtimox は 2010 年 8 月 5 日，Food and Drug Administration（FDA）からシャーガス病の治療薬として希少疾病用医薬品指定を受けた。また，nifurtimox は世界保健機関（WHO）によって生命維持に重要な医薬品に分類されており，WHO のシャーガス病必須医薬品リストに掲載されている。

外部対照を用いた検証的試験について[1]

nifurtimox の臨床開発の検証段階では第 III 相試験が実施され，プラセボ対照試験デザインを用いることは非倫理的であるためヒストリカルコントロール群が設定された。第 III 相試験は 2 つの Part から構成され，Part 1 はシャーガス病と診断された小児の nifurtimox の治療，及び 1 年間の追跡調査から成り，Part 2 は FDA の要請により Part 1 に登録された被験者を対象に part 1 完了後更に 3 年間の追跡調査で計画された。nifurtimox は Part 1 の結果に基づき 2020 年 8 月 6 日に FDA より迅速承認され，2021 年 7 月時点で Part 2 は結果の公表には至っていない（日本では承認されていない）。以降，主に Part 1 の試験デザイン，及び迅速承認の審査過程での主要評価項目の解析部分の議論について触れる。Part 1 の試験デザインはヒストリカルコントロール，年齢による層別ランダム化二重盲検並行群間比較試験で，試験デザインの構成要素は以下のとおりである。

- 対象集団:0 歳から 18 歳未満のシャーガス病患者

- 主要評価項目：以下の 2 つの ELISA 法による血清学的アッセイに基づく治療終了後 12 ヵ月の sero-reduction と sero-conversion による複合評価に基づく治癒，未治癒の二値変数
    - sero-reduction，sero-conversion のどちらか一方で治癒と判定された場合に複合評価は治癒と判定，ただし年齢が 8 ヵ月未満の場合は，sero-conversion のみで治癒を判定
        - sero-reduction は年齢が 8 ヵ月以上 18 歳未満の被験者を対象に lysate-ELISA による光学濃度の減少率と recombinant-ELISA による光学濃度の減少率の平均が 20%以上減少した場合，治癒と判定
        - sero-conversion は全被験者を対象に lysate-ELISA，recombinant-ELISA の両方で免疫グロブリン G（IgG）濃度が陰性の場合，治癒と判定
- 治療[注 1)]：①nifurtimox 錠 60 日間 3 回/日投与群（以下，nifurtimox 60 日間投与レジメン群），②nifurtimox 錠 30 日間 3 回/日投与後プラセボ錠 30 日間 3 回/日投与群（以下，nifurtimox 30 日間投与レジメン群）[注 2)]，及び③ヒストリカルコントロール群

  注 1）割付は年齢層（生後 0～27 日，生後 28 日～8 ヵ月未満，8 ヵ月～2 歳未満，2～18 歳未満）別に nifurtimox 60 日間投与レジメン群と nifurtimox 30 日間投与レジメン群に 2:1 の割合で割り付ける

  注 2）②nifurtimox 30 日間投与レジメン群の投与レジメンは申請投与レジメンではない
- 主要評価項目の主要解析：nifurtimox 錠 60 日間投与レジメン群の治癒割合及び両側 95%信頼区間を算出し，その両側 95%信頼区間の下限値がヒストリカルコントロール群の情報から設定された閾値 16%[注 3)]を超えた場合に nifurtimox 錠 60 日間投与レジメン群のヒストリカルコントロール群に対する優越性が示されたとして評価

  注 3）閾値 16%の設定に用いられたヒストリカルコントロール群の情報として 2 つの文献が参照された。de Andrade et al 1996[1]で報告されたプラセボ群の治癒割合（両側 95%信頼区間）は 5%（1%，13%）であり Sosa et al 1998[1]は 5%（1%，16%）であることからプラセボ群の両側 95%信頼区間の上限値に基づき閾値 16%を設定
- 主要評価項目の副次解析：nifurtimox 錠 60 日間投与レジメン群の治癒割合と nifurtimox 30 日間投与レジメン群の治癒割合の差及び両側 95%信頼区間を算出し，両側 95%信頼区間が 0 を含まない場合に nifurtimox 錠 60 日間投与レジメン群の治療効果と nifurtimox 30 日間投与レジメン群の治療効果に差があるとして評価
- 目標症例数[注 4)]：390 名（nifurtimox 60 日間投与レジメン群：260 名，nifurtimox 30 日間投与レジメン群：130 名）

  注 4）症例数の設定根拠として，nifurtimox 60 日間投与レジメン群に割り付けられる症例数が 260 名の場合，治療終了後 12 ヵ月の nifurtimox 60 日間投与レジメン群の治癒

率を 55%と仮定すると治癒率の 95%信頼区間の下限値が閾値 16%を上回る確率は 99%となる。また，層別割付に用いた年齢層（注 1）参照）の各層に対して最低目標症例数として 38 例を設定

主要解析の結果，及び FDA のレビュー結果

主要解析の結果，nifurtimox 60 日間投与レジメン群の治癒率（両側 95%信頼区間）は 32.9%（26.4%，39.3%）であり信頼区間の下限値 26.4%は閾値 16%を超えているため，申請者は優越性は検証されたと結論付けた（表 3-1-5-2-1）。

しかし FDA の Statistical Review は主要評価項目，及びヒストリカルコントロールに基づく閾値設定について否定的なコメントを出してる[2]。

主要評価項目について，以下の 4 点に触れている。

- ELISA 法による血清学的アッセイは定量的検査よりはむしろ定性的検査としてデザインされているため sero-reduction（光学濃度の 20%以上の減少）の臨床的意義が sero-conversion（IgG 濃度陰性）ほど明確ではない
- 光学濃度と抗体力価の関係が線形でないため 20%の閾値に基づいた sero-reduction による評価が困難である
- 層別割付の年齢層ごとの目標症例数が満たされず生後 0～27 日，生後 28 日～8 ヵ月未満の層の症例が少なくなり，結果的に複合評価の治癒例は，先に評価上の問題点に触れた sero-reduction の治癒達成例に大きく依存してしまった
- lysate-ELISA から得られた光学濃度と recombinant-ELISA から得られた光学濃度の関係を分析した結果，これらの平均により sero-reduction を評価すべきではない

一方で，lysate-ELISA と類似した ELISA を用いた先行試験（Sosa et al.1998，ヒストリカルコントロールとして参照された試験）の成績より，12 か月後の光学濃度の 20%以上の減少が 24 か月後，及び 48 か月後の sero-conversion と相関していることを挙げ，sero-reduction として用いられた光学濃度の 20%以上の減少に対して肯定的なコメントも出している。

次に，ヒストリカルコントロールに基づく閾値設定の適切性について，限られた情報から閾値 16%が設定されていることに触れ，ヒストリカルコントロール群との比較よりは同時対照である nifurtimox 30 日間投与レジメン群との比較で評価すべきとしている。

以上より，FDA は lysate-ELISA と recombinant-ELISA を分けて主要評価項目の副次解析の再解析を実施し，どちらの測定からも nifurtimox 錠 60 日間投与レジメン群と nifurtimox 30 日間投与レジメン群を比較して用量反応効果がある結果を得ている（表 3-1-5-2-2）。

最終的に FDA は，nifurtimox 錠 60 日間投与レジメン群と nifurtimox 30 日間投与レジメン群を比較して用量反応効果が得られたこと，副次評価項目の解析からも nifurtimox 錠 60 日間投与レジメン群の有効性を支持する結果が得られたこと，及び Part 2 の長期追跡による更なる有効性の検証を条件に，Part 1 の段階で迅速承認したようである。先にも述べた通り 2021 年 7 月の時点で，

Part 2 は結果の公表には至っていない。

申請者の治験実施計画書[1]からは一定の根拠に基づき Part 1 の主要評価項目を sero-reduction と sero-conversion の複合評価項目として設定したことが読み取れる。sero-reduction との複合評価項目にした理由の 1 つとして，小児，及び青年のシャーガス病患者は sero-conversion が治療から数年後に認められる場合があり長期の追跡が必要となるため sero-reduction を sero-conversion の代替変数として位置付けたことが伺える。なお，Part 2 の主要評価項目は治療終了後 4 年の sero-conversion である。

まとめ

nifurtimox の検証的試験は，希少疾病用医薬品の臨床開発で見られる 2 つの特徴を持っていた。1 つはプラセボ投与が倫理的に許容されないことによるヒストリカルコントロール群の設定，もう 1 つは FDA の迅速承認制度の活用である。希少疾病用医薬品開発では，ヒストリカルコントロールとして参照できるデータは限られている場合が多く，今回のような閾値の妥当性という点では多くの場合困難が伴うと考えられる。加えて迅速承認制度の活用と関連した Part 1 の主要評価項目として設定した複合評価項目を構成する 2 つの変数のうち 1 つの変数は代替変数の要素を持っていたが，当該変数の臨床的意義について審査過程で懸念が示されたため，ヒストリカルコントロールとの比較が困難となった。

今回，本ケースをヒストリカルコントロールとの比較に基づく薬効評価が困難となった事例として紹介した。このようなリスクは完全には取り除くことはできないが，リスクを小さくするための規制当局との事前協議は希少疾病用医薬の開発では特に重要であると考えられる。

表 3-1-5-2-1 主要評価項目の主要解析の結果

解析対象集団	解析項目	30-Day	60-Day
FAS	N	111	219
	Total cure (rate)	21 (18.9)	72 (32.9)
	Serocnversion	5 (4.5)	10 (4.6)
	≧20% decrease in optical density	16 (14.4)	62 (28.3)
	No cure	90 (81.1)	147 (67.1)
	95% CI for cure rate	(11.2, 26.7)	(26.4, 39.3)

注釈）30-Day：nifurtimox 30 日間投与レジメン群，60-Day：nifurtimox 60 日間投与レジメン群
FDA MULTI-DISCIPLINE REVIEW[2]をもとに作成

表 3-1-5-2-2 FDA が実施した主要評価項目の副次解析の再解析結果

解析項目	Lysate-ELISA		Recombinant-ELISA	
	30-Day	60-Day	30-Day	60-Day
	N=111	N=219	N=111	N=219
Total cure (rate)	21 (18.9)	70 (32.0)	24 (21.6)	76 (34.7)
Serocnversion	6 (5.4)	11 (5.0)	7 (6.3)	11 (5.0)
≧20% decrease in optical density	15 (13.5)	59 (26.9)	17(15.3)	65 (29.7)
Cure RD of 60-day vs. 30-day (95% CI)	13.0 (3.5, 22.6)		13.1 (3.2, 23.0)	
p-value to test the difference between groups	0.007		0.010	

注釈）RD: rate difference, 30-Day：nifurtimox 30 日間投与レジメン群，60-Day：nifurtimox 60 日間投与レジメン群
FDA MULTI-DISCIPLINE REVIEW[2]をもとに作成

参考文献

[1] ClinicalTrials.gov. "Prospective Study of a Pediatric Nifurtimox Formulation for Chagas' Disease." NCT02625974 https://clinicaltrials.gov/ct2/show/NCT02625974

[2] Food and Drug Administration. “nifurtimox (LAMPIT) MULTI-DISCIPLINE REVIEW.” (2018). https://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/213464Orig1s000MultidisciplineR.pdf

3.1.5.3　eteplirsen（Exondys 51™ Injection）

eteplirsen と筋ジストロフィーについて

筋ジストロフィーは，骨格筋繊維の変性・壊死と不完全再生のサイクルを繰り返しながら間質の繊維化・脂肪化が進行する遺伝性疾患群である。臨床的には進行性の骨格筋萎縮と筋力低下を呈し，日常生活動作が低下するだけでなく，呼吸筋不全や心不全といった重篤な合併症を併発することもある難病である[1]。

筋ジストロフィーは，遺伝形式や遺伝子産物によって，様々なサブタイプに分類される。デュシェンヌ型筋ジストロフィーは，ジストロフィンと呼ばれる細胞膜裏打ち蛋白質の欠損によって発症しており，男児 3500 人に 1 人が発症するといわれている。運動発達の遅れによって 2～3 歳に病気と診断され，12 歳までには車椅子での生活となり，14 歳を過ぎると 1/3 の患者に心筋障害を生じる[1]。

eteplirsen はエクソン・スキップ治療と呼ばれる治療メカニズムに基づく治療薬であり，ジストロフィン蛋白発現を回復させることによって，デュシェンヌ型筋ジストロフィーの症状悪化を緩和させる薬剤である[1]。eteplirsen の申請時点で，デュシェンヌ型筋ジストロフィーに対して FDA に承認されている薬剤はなかった。eteplirsen は，FDA によって 2016 年に迅速審査によって承認され，優先審査，及び希少疾病薬の指定も受けた[2]。なお，eteplirsen は日本では承認されていない。

eteplirsen の臨床試験：201 試験，202 試験

eteplirsen の臨床試験としては，201 試験が唯一のランダム化二重盲検プラセボ対照試験である。201 試験は米国の 1 施設で実施され，12 名のデュシェンヌ型筋ジストロフィー患者が登録された。201 試験の終了後，12 名の被験者全員が非盲検の継続長期投与試験（202 試験）に移行し，eteplirsen の投与が継続された。以降の記載は FDA の Statistical Review and Evaluation からの内容の紹介である[3]。

201 試験は，ランダム化二重盲検プラセボ対照試験であり，適格性が確認された被験者は，50mg/kg 群，30mg/kg 群，プラセボ群のいずれかにランダム化され，28 週間の投与が行われた（図 3-1-5-3-1）。201 試験の主要評価項目は，筋肉生検組織中のジストロフィン陽性繊維（dystrophin positive fiber）の割合についての，ベースラインからの変化量である。デュシェンヌ型筋ジストロフィーはジストロフィン蛋白の欠損によって発症することが知られているため，201 試験の主要評価項目は薬力学的な代替評価項目である。症例数は全体で 12 例（4 例/群）である。なお，症例数については，有効性解析に関する検出力に基づいた設定はされていない。

201 試験の主要評価項目（筋肉生検組織中のジストロフィン陽性繊維の割合）の結果を表 3-1-5-3-1 に示す。投与前後差について，30 mg/kg 群では 22.95%の増加が認められた一方で，50 mg/kg 群では 0.79%の増加しか認められなかった。

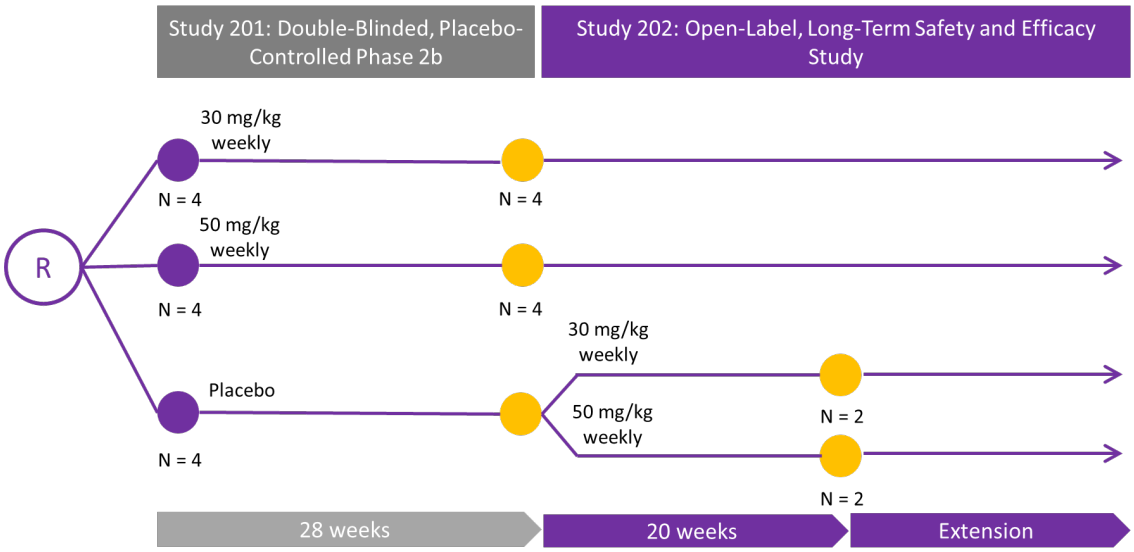


図 3-1-5-3-1　　eteplirsen の臨床試験（201 試験，及び 202 試験）の概要

FDA STATISTICAL REVIEW [3] Figure 1 をもとに作成

表 3-1-5-3-1　　筋肉生検組織中のジストロフィン陽性繊維の解析結果（201 試験）

時点		プラセボ群 N = 4	30 mg/kg 群 N = 4	50 mg/kg 群 N = 4
Baseline	Mean	15.64	18.19	11.00
	Median	15.58	17.80	11.51
	SD (SE)	10.742 (5.371)	5.501 (2.751)	4.668 (2.334)
	Min, Max	3.2, 28.2	11.9, 25.3	5.4, 15.6
On-Treatment	Mean	11.59	41.14	11.79
	Median	9.44	38.77	11.81
	SD (SE)	7.130 (3.565)	10.097 (5.049)	4.456 (2.228)
	Min, Max	5.7, 21.7	32.7, 54.3	6.4, 17.2
Change from Baseline	Mean	-4.05	22.95	0.79
	Median	-6.13	23.46	2.52
	SD (SE)	5.834 (2.917)	5.792 (2.896)	7.099 (3.549)
	Min, Max	-8.5, 4.5	15.9, 29.0	-9.3, 7.4
	p-value*		0.002	0.958

* 共分散分析モデルに基づく。モデルにおいては，順序データとしての投与群（プラセボ群，30mg/kg 群，50mg/kg 群）と固定効果とし，ベースライン値とデュシェンヌ型筋ジストロフィーと診断されてからの期間を共変量とした。

FDA STATISTICAL REVIEW [3] Table 3 をもとに作成

202 試験は，201 試験に参加した被験者を対象とした継続長期投与試験であり，eteplirsen 50mg/kg または 30mg/kg が投与される非盲検試験である。投与期間は約 140 週間である。202 試験の主要評価項目としては，臨床的な評価項目である 6 分間歩行試験が採用されている。また，ヒストリカルコントロール群が 2 つの患者レジストリを用いて作成されており，単群で非盲検の継続長期投与試験であるものの，ヒストリカルコントロール群との比較により臨床的な主要評価項目の改善を示す試みが行われた。eteplirsen 投与例とヒストリカルコントロールとの比較に関する統計解析手法については，Statistical Review and Evaluation に明記されていなかった。なお，eteplirsen のヒストリカルコントロール群との比較に関する文献[4]では，ベースラインの値を共変量とした共分散分析が適用されていた。

202 試験で用いられたヒストリカルコントロールについて

202 試験でのヒストリカルコントロール群について，申請者は，eteplirsen 投与例の観察後に，ヒストリカルデータを入手した。ヒストリカルデータは 2 つの患者レジストリから作成された。eteplirsen 投与例とマッチさせるために，以下の条件で患者の選択を行った。

1. ベースライン時点でのコルチコステロイドの使用
2. 6 分間歩行試験に関する十分な経時データの存在
3. 7 歳以上の年齢
4. あらゆるエクソン・スキップ治療が適用できる遺伝子型
5. エクソン 51 に関するスキップ治療が適用できる遺伝子型

患者選択の結果，186 名のレジストリデータから，13 名の患者がヒストリカルコントロール群として選択された。

ヒストリカルコントロール群の特定後に，申請者は 6 分間歩行試験に関する有効性の解析を実施した（表 3-1-5-3-2）。eteplirsen 投与例のベースライン値は 363.2m であり，36 か月後の値は 256.4m であった。ヒストリカルコントロール群のベースライン値は 357.6m であり，36 か月後の値は 115.1m であった。LS Mean の群間差は 141m であり，群間差の検定の p 値は 0.009 であった。なお，Statistical Review and Evaluation を確認する限り，上記の 5 つの基準で被験者集団を同等にする試みがされているが，傾向スコアなどによる被験者レベルでのマッチングは実施されていない。

表 3-1-5-3-2　　ヒストリカルコントロールを用いた 6 分間歩行試験の解析結果（202 試験）

群		年齢	6 分間歩行試験（ベースライン）	6 分間歩行試験（36 か月後）**
ヒストリカルコントロール	N	13	13	13
	Mean / LS Mean* (SE)	9.45 (0.403)	357.6 (18.51)	115.1 (33.54)
	Min, Max	7.3, 11.8	200, 458	
eteplirsen	N	12	12	12
	Mean / LS Mean* (SE)	9.41 (0.342)	363.2 (12.18)	256.4 (33.11)
	Min, Max	7.3, 11.0	256, 416	

* LS Mean は 6 分間歩行試験（36 か月後）に対して算出されている。

** LS Mean の群間差（36 か月後）は 141m であり，p 値は 0.009 である。

FDA STATISTICAL REVIEW [3] Table 5 をもとに作成

ヒストリカルコントロールに関する FDA のレビュー結果

本事例であるが，FDA の Statistical Review and Evaluation では，「ヒストリカルコントロールを用いた非盲検の継続長期投与試験は，統計学的に解釈できない」という厳しい評価を受けている。Statistical Review and Evaluation では，以下の点が指摘されている。

- ICH E10 ガイドライン「臨床試験における対照群の選択とそれに関連する諸問題」[5]では「外部対照試験を採用すべきか考慮するのは，一般に，被験治療が全ての既存の治療法より優れているとの事前の確信がきわめて強い…」と記載があるが，この条件は eteplirsen には該当しない。
- ICH E10 ガイドラインでは「外部対照を採用するのは，エンドポイントが客観的であり」との記載があるが，6 分間歩行試験はモチベーションによる影響を受ける。
- ヒストリカルコントロールの受け入れについて，Pocock [6] は「治療の評価方法が新規試験と同じである」，及び「新規試験とほぼ同じ組織で実施されている」という基準を挙げている（3.1.2 節参照）。これは，6 分間歩行試験のようなエンドポイントの評価においては特に重要であるが，この要件は満たされていない。
- 更に重要な点として，今回の申請おいて，ヒストリカルコントロール群は事後的に特定された点が挙げられる。そのため，定量化できない選択バイアスがもたらされた可能性がある。もしヒストリカルコントロールを用いるのであれば，対照群の選定や，選択基準の設定は，実薬群と対照群のアウトカムを知らない状態で事前に計画されるべきであった。また，ヒストリカルコントロールを用いた 202 試験の成功基準は，予め治験実施計画書で規定されていなかった。

まとめ

ヒストリカルコントロールの利用に関して否定的な見解であった事例として，eteplirsenの事例を取り上げた。なお，ヒストリカルコントロール群が設定された 202 試験は継続長期投与試験であるため，ヒストリカルコントロールの利用に関する否定的な見解については直接的な承認の妨げにはならなかったようである。ただし，迅速承認に関して，長期間（2 年間）の有効性を確認するためのランダム化比較試験を実施するという条件が付与された。FDA の Statistical Reviewer の見解については，審査上で論点になりうる点を理解する上での参考事例となれば幸いである。特に，ヒストリカルコントロールの利用に関しては，事前に検討し治験実施計画書に十分に記載する点は重要であり，必要に応じて規制当局への相談も重要と考える。

参考文献

[1] 戸田達史. "2. 筋ジストロフィーの分子機構と治療戦略." 日本内科学会雑誌 105.9 (2016): 1578-1587.

[2] ミクス Online. ”米 FDA 世界初の DMD 治療薬 Exondys51 を承認.” https://www.mixonline.jp/tabid55.html?artid=54633.

[3] Food and Drug Administration. “EXONDYS 51™ (eteplirsen) Statistical Review and Evaluation.” (2016). https://www.accessdata.fda.gov/drugsatfda_docs/nda/2016/206488Orig1s000StatR.pdf

[4] Mendell, Jerry R., et al. "Eteplirsen Study Group and Telethon Foundation DMD Italian Network. Longitudinal effect of eteplirsen versus historical control on ambulation in Duchenne muscular dystrophy." Ann. Neurol 79.2 (2016): 257-271.

[5] 厚生労働省. “臨床試験における対照群の選択とそれに関連する諸問題.” (2001).

[6] Pocock, Stuart J. "The combination of randomized and historical controls in clinical trials." Journal of chronic diseases 29.3 (1976): 175-188.

3.1.5.4　ビルトラルセン（ビルテプソ®）

筋ジストロフィーについて

eteplirsen の事例を参照。

ビルトラルセンについて

ビルトラルセンはデュシェンヌ型筋ジストロフィー（DMD）の原因遺伝子であるジストロフィンのエクソン 53 を標的とするアンチセンス核酸であり，エクソン 53 を取り除く（スキップする）ことでアミノ酸読み取り枠を回復させる薬剤である。これにより，正常よりも短いが両端の構造を保持したジストロフィンタンパク質を発現させることにより作用を示すと考えられる[1]。また，ビルトラルセンは条件付き早期承認制度の適用を受け，以下を承認条件として 2020 年 3 月 25 日に日本で製造販売が承認された薬剤である[2]。

1. 本剤の安全性，及び有効性に関するデータを早期に収集すること
2. 国内での治験症例が極めて限られていることから，再審査期間中は，全症例を対象とした使用成績調査を実施することにより，本剤の使用患者の背景情報を把握するとともに，本剤の安全性，及び有効性に関するデータを早期に収集し，本剤の適正使用に必要な措置を講じること。
3. 本剤の有効性，及び安全性の確認を目的とした臨床試験，及び国内レジストリを用いた調査を実施し，終了後速やかに試験成績，及び解析結果を提出すること。

後述するように，ビルトラルセンの審査において海外レジストリデータをヒストリカルコントロールとして利用していたが，承認後に国内レジストリを用いた調査を実施するよう指示を受けている。また，有効性と安全性の確認を目的としたプラセボ対照ランダム化比較第 III 相試験が現在実施中である[3]。また，ビルトラルセンは米国においても 2020 年 8 月 13 日に販売承認を受けている[4]。

他で承認されている薬剤の有無（申請時）

他の治療薬として，米国においてビルトラルセンと同じモルフォリノ核酸である eteplirsen がエクソン 51 スキッピングで治療可能な DMD を適応症として 2016 年 9 月に，また，golodirsen がエクソン 53 スキッピングで治療可能な DMD を適応症として 2019 年 12 月に迅速承認されている。一方，欧州においてナンセンス変異のリードスルーを誘導する低分子医薬品である ataluren が，ナンセンス変異型 DMD を適応症として 2014 年 8 月に条件付き承認されている。なお，海外の診療ガイドラインにおいて，これらの薬剤は申請時点では記載されていない。また，これらの薬剤は申請時点で国内において承認されていない。

外部対照を利用した試験：201 試験

臨床データパッケージの 3 試験のうち 201 試験において外部対照（自然歴：海外レジストリ）との比較が行われていた。201 試験は，DMD の 4 歳から 10 歳未満の男児 16 名を対象として海外 6

施設（米国，及びカナダ）で実施した用量設定試験であり，ビルトラルセン 40 または 80 mg/kg を週 1 回，20 または 24 週間投与した際の有効性，安全性，及び薬物動態を評価した。本試験は，安全性を確認するためのプラセボを対照とした 4 週間のランダム化二重盲検期と，それに続く 20 週間の非盲検期の計 24 週間で構成された（図 3-1-5-4-1 参照）。主要評価項目のジストロフィン発現量はベースライン時点と 24 週時点でのみ測定されており，薬剤群とプラセボ群との比較は行われていない。DMD 治療薬の医薬品開発において，特に開発後期の臨床試験の主要評価項目としては 6 分間歩行距離が標準的に使用されてきたものの，プラセボに対する優越性が検証された品目は申請時において存在しない。また，FDA における DMD 治療薬開発のためのガイダンス[5]においても，ジストロフィン異常症（DMD 等）の医薬品開発において，組織レベル等で骨格筋の量を確実に反映するバイオマーカーは，十分な科学的エビデンスと許容可能な分析法で裏付けられる場合には，迅速承認を支持する代替評価項目として有益である可能性がある旨が記載されている。そのため，本試験において申請者は主要評価項目にジストロフィンタンパク発現を用いており，ジストロフィンタンパク発現の臨床的意義を評価するため 6 分間歩行距離といった運動機能も副次評価項目に含め，運動機能検査の比較対照として自然歴研究を用いている。

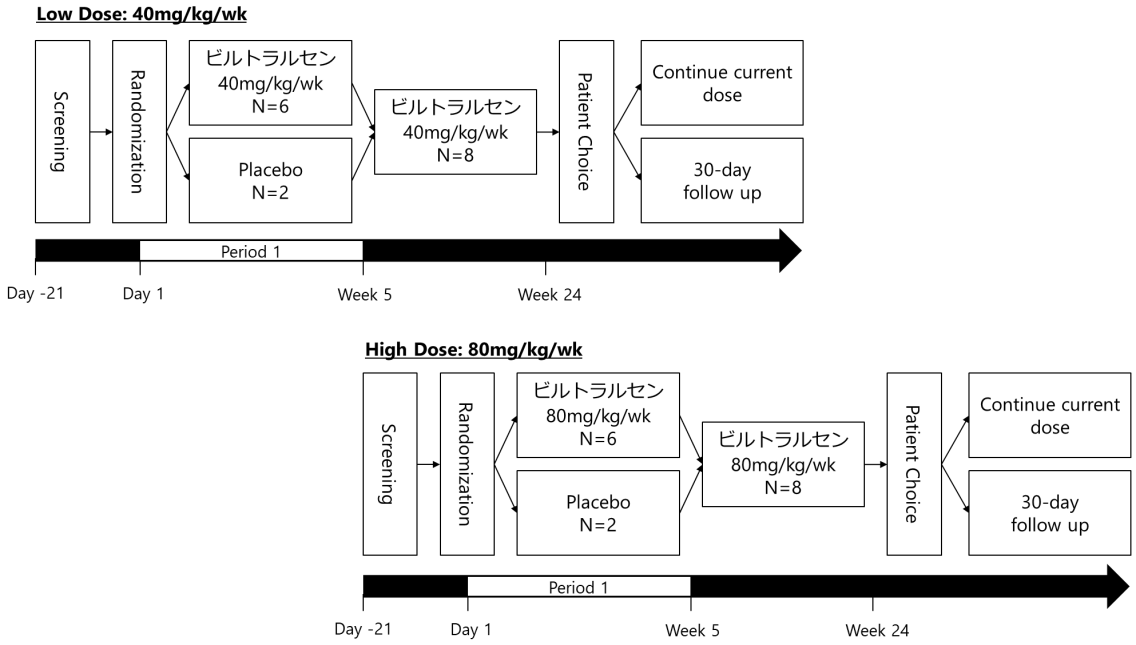


図 3-1-5-4-1：201 試験の概要

ビルテプソ点滴静注 250 mg に関する資料[6]に基づき作成

症例数設定の根拠

申請資料[6]に症例数の設定に関して以下の記載があり（一部省略），主に安全性，有効性に関してどの程度の推定が可能かということに基づいて設定しており，仮説の検証を目的としていない

ことがうかがえる。

発現割合が 10%と考えられる有害事象について，最低 1 件の発現が認められる確率は，N=16 の時に 81%である。

用量群間（各 8 例）では 1.5 SD 以上の差を 80%以上の確率で，1.7SD 以上の差を 90%以上の差で検出可能であると考えた。ジストロフィンの発現（%）ついては，単独の用量群（8 例）において，SD が 4%と仮定した場合に，正常の 5%のジストロフィン増加を検出力 80%で検出することが可能と考えた。

自然歴の比較可能性

審査報告書に以下の記載があり，比較可能性を高めるため事前に CINRG DNHS[1]で収集されたデータに合わせるように 201 試験がデザインされたことがうかがえる。

海外 201 試験の運動機能評価の測定時点，試験実施施設，試験業務の委託先，運動機能評価の手順書，及びトレーニング方法は，CINRG の自然歴集団と比較可能なデザインとなるように設定した。

また，CINRG DNHS の対照として用いるデータは，少なくとも 12 ヵ月の時間機能（有効性）検査データがある，年齢（4 歳以上 10 歳未満），地域（北米）といった条件を満たす対象と設定している。選択・除外基準の設定の参考資料として申請者は 2018 年 2 月に発行された FDA の DMD 治療薬開発のためのガイダンスを挙げている。

その結果，男児の DMD 患者 65 例が基準を満たし，そのうち，9 例はエクソン 53 スキッピング治療対象の患者，56 例はそれ以外の患者であった。201 試験の被験者と自然歴から選択された被験者の背景因子の結果は表 3-1-5-4-1 の通りである。

表 3-1-5-4-1：201 試験と自然歴集団の患者背景の比較

	海外 201 試験			自然歴集団		
	40 mg/kg	80 mg/kg	全体	エクソン 53 スキッピング	エクソン 53 スキッピング以外	全体
評価例数	8	8	16	9	56	65
年齢（歳）	7.5±1.75	7.2±2.03	7.4±1.84	6.3±1.07	7.2±1.36	7.1±1.35
人種（%）						

1 米国の筋ジストロフィーの臨床試験ネットワークである cooperative international neuromuscular research group （CINRG）が実施した Duchenne natural history study（DNHS）と呼ばれる研究のこと

	海外 201 試験			自然歴集団		
	40 mg/kg	80 mg/kg	全体	エクソン 53 スキッピング	エクソン 53 スキッピング以外	全体
白人	8 (100)	7 (87.5)	15 (93.8)	7 (77.8)	48 (85.7)	55 (84.6)
黒人	0	0	0	0	1 (1.8)	1 (1.5)
アジア人	0	1 (12.5)	1 (6.3)	1 (11.1)	3 (5.4)	4 (6.2)
その他	0	0	0	1 (11.1)	4 (7.1)	5 (7.7)
身長(cm)	114.6±6.50	112.2±9.97	113.4±8.22	111.3±7.57	117.0±10.20	116.2±10.03
体重	23.7±4.70	22.3±6.16	23.0±5.34	21.6±3.99	24.4±6.22	24.0±6.02
BMI	17.9±2.28	17.4±2.02	17.7±2.10	17.3±1.99	17.6±2.30	17.5±2.25
立ち上がり時間(rise/秒)	0.26±0.058	0.25±0.090	0.25±0.074	0.23±0.068	0.21±0.092	0.22±0.089
10m 歩行/走行時間(m/秒)	1.67±0.385	1.88±0.357	1.77±0.374	1.92±0.458	1.90±0.470	1.91±0.465
4 段階段昇り時間(task/秒)	0.27±0.081	0.32±0.081	0.30±0.082	0.30±0.082	0.27±0.105	0.28±0.112
NSAA Total Score	24.8±5.92	23.8±5.09	24.3±5.36	28.0±6.68	25.2±5.11	25.7±5.37
6 分間歩行距離(m)	391.4±33.27	353.4±106.32	372.4±78.59	428.4±63.50	403.2±184.51	408.0±167.16

ビルテプソ点滴静注 250 mg, 審査報告書[2]の表 40 をもとに作成

平均値±標準偏差

有効性(ジストロフィン発現)の比較

主要評価項目である筋生検によるジストロフィンタンパク発現のベースラインからの変化は表 3-1-5-4-2 のとおりであり, ミオシン重鎖及びα-アクチニンのいずれで標準化した場合においても, ビルトラルセン 40 及び 80 mg/kg 群ともに, 24 週投与（25 週時）後のジストロフィン発現はベースライン時と比較して統計学的に有意に増加した（対応のある t 検定, 多重性の調整なし)。

表 3-1-5-4-2:ウェスタンブロット法によるジストロフィン発現のベースラインからの変化

投与群	40 mg/kg 群	80 mg/kg 群
評価例数	8	8
ミオシン重鎖による標準化（%）		
ベースライン	0.3 ± 0.10	0.6 ± 0.82
25 週時	5.7 ± 2.37	5.9 ± 4.50
ベースラインからの変化	5.4 ± 2.40	5.3 ± 4.48
P 値	0.0004	0.0123
α-アクチニンによる標準化（%）		
ベースライン	0.2 ± 0.22	0.4 ± 0.67

投与群	40 mg/kg 群	80 mg/kg 群
25 週時	5.4 ± 2.79	3.7 ± 2.37
ベースラインからの変化	5.2 ± 2.83	3.3 ± 2.47
P 値	0.0012	0.0074

ビルテプソ点滴静注 250 mg，審査報告書[2]の表 29 をもとに作成

平均値±標準偏差

P 値は対応のある t 検定を用いて算出した

有効性（運動機能検査）の比較

ビルトラルセン（40 mg/kg，80 mg/kg）の治療を受けた被験者 16 例と CINRG DNHS の自然歴群（DNHS 群）65 例における，25 週時のベースラインからの運動機能検査の変化を比較した結果は表 3-1-5-4-3 の通りである。その結果，10m 歩行/走行時間（m/秒），10m 歩行/走行時間（秒），立ち上がり時間（秒）及び 6 分間歩行距離では，ビルトラルセン群の方が有意に改善していた（P=0.0029，P=0.0462，P=0.0366，P=0.0471）。運動機能検査の結果で改善が認められたことにより，ビルトラルセン投与により誘発されるジストロフィンタンパク発現は臨床的ベネフィットをもたらす可能性が示唆された旨，記載されている。

表 3-1-5-4-3：時間機能検査 25 週時におけるベースラインからの変化

	海外 201 試験			自然歴集団		
	40 mg/kg	80 mg/kg	全体	エクソン 53 スキッピング	エクソン 53 スキッピング以外	全体
評価例数	8	8	16	9	56	65
10m 歩行/走行時間（m/秒）						
ベースライン	1.67±0.385	1.88±0.357	1.77±0.374	1.92±0.458	1.90±0.470	1.91±0.465
変化量（25 週）	0.21±0.291	0.24±0.222	0.23±0.251	0.08±0.141	-0.05±0.336	-0.04±0.327
P 値	0.0365	0.0132	0.0029			
10m 歩行/走行時間（秒）						
ベースライン	6.30±1.587	5.55±1.336	5.93±1.469	5.45±1.173	5.64±1.745	5.61±1.671
変化量（25 週）	-0.65±1.225	-0.66±0.921	-0.66±1.047	-0.23±0.425	0.10±1.462	0.08±1.414
P 値	0.2066	0.0789	0.0462			
立ち上がり時間（秒）						
ベースライン	4.17±1.146	4.76±2.580	4.44±1.956	4.61±1.424	5.70±3.210	5.55±3.041
変化量（25 週）	0.05±1.446	-0.44±0.750	-0.19±1.141	-0.92±0.402	0.78±1.859	0.66±1.845
P 値	0.2077	0.0543	0.0366			
6 分間歩行距離（m）						

	海外 201 試験			自然歴集団		
	40 mg/kg	80 mg/kg	全体	エクソン 53 スキッピング	エクソン 53 スキッピング以外	全体
ベースライン	391.4±33.27	353.4±106.32	372.4±78.59	428.4±63.50	403.2±184.51	408.0±167.16
変化量(25 週)	15.6±26.40	44.0±41.98	28.9±36.31	70.7±97.16	-90.0±162.51	-65.3±162.60
P 値	0.1178	0.0823	0.0471			

ビルテプソ点滴静注 250 mg に関する資料[6]の表 2.7.6.1.4.2-6 ～ 9 をもとに作成

平均値±標準偏差

P 値はビルトラルセン vs 自然歴集団(全体 N=65)を比較した結果である。また, P 値は mixed model for repeated measures を用いて算出した。治療, 来院週, 来院週と治療の交互作用を因子とし, ベースライン時の年齢及びベースライン値を共変量とし, 分散共分散構造には無構造を用いた。Kenward-Roger 近似を用いて分母の自由度を推定した。

PMDA の外部対照への言及

審査報告書で外部対照に対して以下のように評価されている。

海外 201 試験と CINRG の自然歴集団等の外部対照との比較について, ランダム化により比較可能性の担保された集団間での比較ではなく, 探索的な検討ではあるものの, 得られた試験成績から, 本剤投与によるジストロフィン発現増加により, 運動機能が改善する傾向は示唆されている。また, 国内第 I/II 相試験, 及び海外 201 試験において, いずれもジストロフィン発現の増加が認められ, その発現量の意義についても一定の説明はなされている。これらの点を踏まえると, 本剤投与により運動機能に対する有効性は期待できる。

条件付き早期承認制度の利用

申請者は審査報告書において, 以下の理由(一部抜粋)から条件付き早期承認制度の利用が適切である旨の説明をしており, 外部対照との比較結果が良好であったことが理由の一つとして挙げられている。

- 国内外の臨床試験において, ビルトラルセン投与によりジストロフィンタンパク増加が認められたこと, 海外 201 試験の本剤群では外部対照である CINRG の自然歴集団と比較して, 運動機能の改善傾向が認められていることから, 本剤の一定の有効性が示されたと考えられる。
- 運動機能等, 臨床的な評価項目に基づきビルトラルセンの有効性を検証することを目的とした臨床試験を実施する場合, 国際共同治験として実施した場合でも被験者の組入れに〇年程度, 試験全体としては 5 年程度を要する(〇は審査報告書中, 黒塗り)。
- 以上より, 本剤の対象疾患は希少かつ重篤であり, 有効な治療法も少なこと等から, 現在の

臨床データパッケージに基づき，条件付き早期承認制度を利用して承認申請することが適切と考えた。

PMDA はこの説明を了承している。

まとめ

本節では計画時に試験で収集する項目をレジストリの収集項目に合わせた例を紹介した。ランダム化により比較可能性の担保された集団間での比較ではないことが留意点として挙げられているものの，外部対照との比較は運動機能の改善を示すエビデンスとして PMDA から一定の評価を受けており，これが条件付き早期承認を受ける 1 つの要因になっていることがうかがえる。このことから，事情により対照が置けないような領域における薬剤開発で，試験計画の段階で外部対照に測定時点や実施施設などを合わせ，比較可能性を向上させる工夫をすることが有効な一つの選択肢であることが分かる。

参考文献

[1] 日本新薬株式会社. ビルテプソ製品情報概要 (2020). https://med.nippon-shinyaku.co.jp/viltepso/product/pharmacology.html

[2] 医薬品医療機器総合機構. “ビルテプソ点滴静注 250 mg，審査報告書.” (2020). https://www.pmda.go.jp/drugs/2020/P20200408002/530263000_30200AMX00428_A100_1.pdf

[3] ClinicalTrials.gov. "Study to Assess the Efficacy and Safety of Viltolarsen in Ambulant Boys With DMD (RACER53)." NCT04060199. https://clinicaltrials.gov/ct2/show/NCT04060199

[4] 日本新薬株式会社. デュシェンヌ型筋ジストロフィー治療剤「ビルテプソ」の米国 FDA 承認取得に関するお知らせ (2020). https://www.nippon-shinyaku.co.jp/file/download.php?file_id=5503

[5] Food and Drug Administration. "Duchenne Muscular Dystrophy and Related Dystrophinopathies: Developing Drugs for Treatment - Guidance for Industry." (2018). https://www.fda.gov/media/92233/download

[6] 日本新薬株式会社. ビルテプソ点滴静注 250 mg に関する資料 (2019) https://www.pmda.go.jp/drugs/2020/P20200408002/index.html

3.2 アダプティブデザイン

3.2.1 概要

アダプティブデザインは，近年，臨床試験の中でも特に研究されてきた方法論の一つであり，新薬開発の期間やコストの抑制，成功確率の上昇などが期待される。また，希少疾患では臨床試験に登録できる症例数に限界があることから，このような革新的な方法論に基づき新規治療の薬効についての十分なエビデンスを効率的に獲得することが期待される。

アダプティブデザインの定義には諸説あるが，例えば，PhRMA Adaptive Designs Working Group [1]では，「蓄積された試験のデータに基づき，Validity（妥当性）とインテグリティ（完全性）を損なうことなく，試験の特性をどのように変更するかを決定する」試験デザインとしている。同白書では，アダプティブデザインの目的は，試験データから学習することと，学習したことを可能な限り迅速に適応させることとしており，また，事後的（ad-hoc）ではなく事前の計画に従って試験デザインは変更されるものであって，不適切なデザインや計画に対する救済ではないとしている。FDA ガイダンス Adaptive designs for clinical trials of drugs and biologics-guidance for industry [2]（以下の邦訳は JPMA によるアダプティブデザインに関する FDA ガイダンスの邦訳[3] から抜粋）では，「臨床試験に参加した被験者の蓄積されたデータに基づいて，試験デザインの 1 つ以上の側面について，予め計画された変更を行うことができる臨床試験デザインと定義する。」とされている。Lesaffre et al. [4] はこの FDA ガイダンスの定義の重要な点として，

（1）事前に計画されたものであること（いつ，なにをトリガーとして変更を実施するか，試験の開始前に定めること），

（2）1 つないしはそれ以上の特定された試験デザインの特徴や仮説に関する変更であること（試験期間中にどのような変更を可能とするのか，試験の開始前に定めること），

（3）その試験に参加した症例から得られたデータに基づくこと（変更は試験から得られたデータに依存し，試験外からの情報に基づいた変更ではないこと）

と解説している。一方，EMA の Reflection Paper on Methodological Issues in Confirmatory Clinical Trials with Flexible Design and Analysis Plan [5]によると，アダプティブデザインは「統計手法が，中間解析での試験デザインの要素（症例数，割付比，治療群の数）変更を第 1 種の過誤確率を完全に制御しつつ許容するデザイン」と定義されている。

アダプテーションには様々な事項が考えられ，以下にいくつか示す。

- 試験の早期中止（群逐次）

ある時点までに蓄積したデータに基づく中間解析により，治療の有効性又は無益性による早期中止を許容するデザインであり，群逐次法の方法論については，Jennison & Turnbull (1999) [6]; [森川 & 山中 訳 [7]]で詳しく説明されている。蓄積したデータに基づいて中間

解析を実施する理由は，参加する症例を不当なリスクにさらさないという倫理的な観点，臨床試験の迅速化とコスト抑制の観点などが考えられ，中間解析の結果に応じてアダプテーションを可能にする群逐次デザインは有用と考えられる。基本的には，事前に規定した時点までに蓄積したデータに基づいて早期中止を判断するため，中間解析の時点や第 1 種の過誤確率の制御，中止境界の設定を事前に計画検討する。また，試験の早期中止を定量的に判断する方法として，中間解析時点での条件付き検出力を評価する場合もある。中間解析の結果によって，その後に登録される症例の評価にバイアスが生じる恐れがあるため，情報の機密性や中止のルールについては注意深く計画を立てる必要がある。また中間データのモニタリングには，外部の独立なデータモニタリング委員会（DMC）が活用されよう。

- 統計学的な有意性を達成するための症例数の増加（症例数再設定）

臨床的に意味のある効果の大きさやそのばらつきを仮定できれば，有意水準や検出力などを適切に設定した下での症例数の検討が可能になる。しかしながら，それらの仮定した効果の大きさやばらつきが真値と乖離すると，計画した症例数では検出力が極めて過剰になる，もしくは，必要な検出力を確保できず，試験結果の再現性に疑念が生じる可能性が考えられる。また，事前の小規模なパイロット試験の結果を過剰に解釈する可能性もあり，その一つの解決方法として Internal pilot study として試験途中の結果に基づいて症例数を再設定することが挙げられる[6]。また，試験の完全性を担保するために，盲検下で症例数再設定を行う方法が複数提唱されている[8, 9, 10, 11]。一方で，群逐次デザインの枠組みで非盲検下での症例数再設定の方法についても複数検討されており，Chow et. al. (2011) [12] で詳しく説明されている。

- 試験途中での，より効果の高い試験治療への割り付け確率の変更（レスポンスアダプティブランダム化）

単純ランダム化や層別ランダム化といった従来のランダム化と比べ，アダプティブランダム化では特定の時点までに割り付けられた症例のデータに基づいて割り付け確率を変更する。その時点までの治療効果に基づいて割付を行うレスポンスアダプティブランダム化では，より効果の高い治療群に割付けられる可能性が高くなる[12]。ただし，評価結果の入手に時間を要する試験では実装に適さないと考えられる。アダプティブランダム化の留意点としては，試験の結果が早期のデータに影響を受ける可能性のほか，組み入れバイアスの懸念が挙げられる。アダプティブランダム化によって引き起こされる組み入れバイアスは Drift と呼ばれ，割り付け確率の変更後に組み入れられた集団の予後因子の分布は，変更前の被験者集団のそれとは異なる可能性がある[4]。アダプティブランダム化の例として，ECMO 臨床試験[13]は新生児遷延性肺高血圧症を対象に ECMO と従来療法の死亡率を比較

する臨床試験であり，この試験では，「新たに割り付ける被験者を，既に組み入れられた症例の結果が良好であった治療への割り付け確率を増加させる」といったプレイ・ザ・ウィナー規則[14]に従ってランダム化が行われた。同じく ECMO を評価した Ware et. al. (1989) [15]では，一定数の死亡例が集積された時点で単群試験に移行するデザインを用いた。これらのレスポンスアダプティブランダム化では症例をより有効な治療に割り付けることができたため，当該治療群での症例数が増え，試験への症例登録も促進された可能性がある。一方で，割り付けが容易に予測できてしまうため，評価にバイアスが生じたり，試験途中で割り付け確率が変わることで対照群との比較可能性に問題が生じたりする場合も考えられる。

その他，アダプティブデザインについては，書籍や総論が多数出版されているため，そちらを参照されたい[4, 12, 16, 17]。

アダプティブデザインの潜在的な利点としては，統計的な効率，倫理的な配慮，薬剤効果の理解の向上，利害関係者への受容性が考えられ，これらは希少疾病領域で臨床試験を実施するうえで非常に重要な要素であると考えられる。例えば後述する中間解析による用量選択を伴う，アダプティブシームレス第 II/III 相試験は，別々に 2 本の臨床試験を実施する場合と比較して，各試験間に生じる期間（white space）を省略できる効率的なデザインである。さらに，適切な多重性調整の下で，用量選択に用いた第 II 相部分のデータを最終解析にも使用可能となる。一方，アダプティブデザインの限界としては，多くの場合複雑な解析手法を必要とし，そのためシミュレーションによる評価も必要になる場合があること，ある側面における効率の向上は別の側面での効率の低下によって相殺される可能性があること，試験の適切な実施及び完全性維持のための運営上の課題があること，アダプテーションの前後で異なる結果が得られた場合の解釈が困難になることが挙げられる。

希少疾病では試験に登録できる症例数が限られるため，単一のアダプティブシームレス第 II/III 相試験の成績を以って申請を目指す場合もあるだろう。EMA の Reflection Paper では，通常であれば，申請には単一の検証的試験で十分な場合であっても，他にその薬物を支持する根拠がなければ，アダプティブシームレス第 II/III 相試験のみの成績では不十分であろうと記載されている[5]。一方，希少疾病の場合は単一のアダプティブシームレス第 II/III 相試験であっても，その試験が限られた症例数から効率的に情報を取得し示しているのであれば，申請に十分であるとみなされうる旨が記載されている。

以下では，乳児血管腫を対象に，アダプティブデザインに基づく海外第 II/III 相試験が実施され承認をされているプロプラノロールの事例を紹介する[18]。

3.2.2　プロプラノロール（ヘマンジオル®シロップ小児用）

乳児血管腫について

乳児血管腫（IH）は皮膚の表面や内部にできる「赤あざ」の一種で，未熟な毛細血管が増殖して

あらわれる良性の腫瘍である。一般的に生後 1～4 週にあらわれ，大きくなる場合は 1 年以内に急速に大きくなり（増殖期），その 90%以上は，5～7 歳までに数年間かけて赤みは少しずつ消えていくが（退縮期），多くの場合「あと」（瘢痕：はんこん）が残る（消失期）。IH のうち，重要臓器，及び感覚器官に影響を及ぼすような病変は 10％程度である。日本人における IH の発症率は 0.8～1.7％と報告されており，日本における平成 24 年時点での出生数 104 万人とこの発症率から算出された患者数は 8,300~17,700 人と推定されている。IH に対する薬物療法としては，様々なものが試みられてきたものの，プロプラノロールの申請時点で承認されている薬剤は日本ではなかった。

プロプラノロールについて

プロプラノロールは非選択的β受容体遮断薬であり，国内では本態性高血圧などですでに承認されている。血管収縮作用，血管内皮増殖因子（以下，「VEGF」），及び塩基性線維芽細胞増殖因子（bFGF）等の発現抑制を介した血管新生抑制作用等により，特に増殖期の IH に対して効果を示すことが期待されており，日本小児血液・がん学会から医療上の必要性の高い未承認薬・適応外薬検討会議に対して開発要望が提出されていた。なお海外では IH を対象に 2014 年 3 月に米国，2014 年 4 月に欧州で承認されている。日本においては乳児血管腫の適応に関して 2015 年に希少疾病用医薬品指定がされ，2016 年 7 月に承認されている。

プロプラノロールの有効性の評価に用いられた試験は国内第 III 相試験（M703101-01），及び海外第 II/III 相試験（V00400 SB 2 01）であった。M703101-01 試験は 30 例からなる単群試験である。一方 V00400 SB 2 01 試験はアダプティブデザインを用いており，中間解析での用量選択を行う試験デザインであった。本節ではアダプティブデザインを用いた V00400 SB 2 01 試験について主に述べる。参考として国内での申請根拠となった M703101-01 試験についても述べる。

アダプティブデザインを用いた海外第 II/III 相試験（V00400 SB 2 01）

海外第 II/III 相試験（V00400 SB 2 01）は海外で実施された多施設共同ランダム化二重盲検プラセボ対照並行群間比較試験である。被験者は①1 mg/kg/日・3 か月 M 実薬群，②1 mg/kg/日・6 か月実薬群，③3 mg/kg/日・3 か月実薬群，④3 mg/kg/日・6 か月実薬群，⑤プラセボ 6 か月投与群のいずれかに，2:2:2:2:1 の割付比で割付が行われた。なお 3 か月投与群（①，③）では 4～6 か月目はプラセボ投与がされている。主要評価項目はベースラインと比較した 6 か月後における乳児血管腫に対する有効率であり，写真判定において，評価時とベースライン（治療開始日）を比較し，IH の状態を「有効（治癒またはほぼ治癒）」または「無効」の 2 段階で評価した。なお本試験は海外試験であり日本は参加していない。

中間解析

本試験は第 II/III 相試験であり，4 つの実薬群とプラセボの比較を行う第 II 相試験に相当するステージ 1 と，選択された 1，2 用量とプラセボを比較する第 III 相試験に相当するステージ 2 からな

る。ステージ 1 の終了時点であるプラセボ群 20 例，各実薬群 40 例が 6 か月での評価を終了または試験中止した時点で中間解析を実施し，投与レジメンの選択が行われた。中間解析の結果に応じて，データモニタリング委員会は以下の A～C のうちいずれかの勧告を選択する。

A. 安全性上の問題があることまたは効果がないことを理由に，試験を中止する。

B. 試験薬の 4 つの用法・用量の中で最も効果があり，安全性に問題がない用法・用量を 1 つまたは 2 つ選択し，それらの試験薬群とプラセボ群のみ試験を継続する。この場合，プラセボ群，及びステージ 1 の中間解析の結果から選択されたプロプラノロール用量群に対して，1:2 の比率でランダムに割付ける。ランダム化された症例が，プラセボ群はステージ 1 との合計で 50 例，選択されたプロプラノロール用量群ではステージ 1 との合計で 100 例/群になるまで，ステージ 1 に引き続き症例登録する。

C. 選択したプロプラノロール用量群のプラセボ群に対する優越性を検証するための検出力が 80%を下回る場合，目標症例数を再設定（追加）する。

なお本試験はアダプティブデザインの下に行われた検証的な試験であること，1 試験の結果に基づき承認申請を行うことを計画していたことから，より強いエビデンスであること示すために，主要解析において有意水準は片側 0.005 を用いている。

<u>多重性調整</u>

本試験デザインでは中間解析で用量が選択されるため多重性の調整が必要であり，閉検定手順が用いられた。例えば，中間解析で 1 つの用量が選択された場合に，選択された用量で帰無仮説を棄却するためには，8 つの併合積仮説を棄却する必要があった。例えば中間解析の結果，用量・用法群④が選択された場合に，$i(=1,2,3,4)$番目の用量・用法群とプラセボ群との有効率の差に関する帰無仮説を$H_{0,i}: \theta_i \leq 0$とし，用量・用法群④に対応する帰無仮説 $H_{0,4}: \theta_4 \leq 0$を棄却するには，$H_{0,4}$を含む併合積仮説 8 個すべてが棄却されなければならない。具体的には①$H_{1234} = H_{0,1} \cap H_{0,2} \cap H_{0,3} \cap H_{0,4}$,②$H_{124} = H_{0,1} \cap H_{0,2} \cap H_{0,4}$，③$H_{134} = H_{0,1} \cap H_{0,3} \cap H_{0,4}$，④$H_{234} = H_{0,2} \cap H_{0,3} \cap H_{0,4}$，⑤$H_{14} = H_{0,1} \cap H_{0,4}$，⑥$H_{24} = H_{0,2} \cap H_{0,4}$，⑦$H_{34} = H_{0,3} \cap H_{0,4}$，⑧$H_{0,4}$である。それぞれの積仮説に対する p 値は，ステージ 1 のデータに対して Simes 法を用いた場合の積仮説に対する p 値(p)と，ステージ 2 のデータのみで計算された p 値(q)をもとに，以下に示す重み付き逆正規分布による統合検定法[19]を用いて算出している。

$$C(p,q) = 1 - \Phi[w_1\Phi^{-1}(1-p) + w_2\Phi^{-1}(1-q)]$$

ここでΦ^{-1}は，標準正規分布の累積分布関数の逆関数である。なお，w_1，及びw_2はステージ 1 でのプラセボ群，及び試験薬群の症例数をn_{1p}，及び$2n_{1p}$，ステージ 2 でのプラセボ群，及び試験薬群の症例数をn_{2p}，及び$2n_{2p}$とした場合に$w_1^2 = n_{1p}/(n_{1p} + n_{2p})$，$w_2^2 = n_{2p}/(n_{1p} + n_{2p})$で与えられる。全体での p 値は 8 つの併合積仮説の p 値のうち最大のものとされた。なお，本タスクフォースで[18]を精査したところ，各ステージの p 値の算出において，分散を併合しない Wald 型の z 検定が用いられたと思われる。

具体的に，本試験結果に関する原著論文[20]の Supplemenral Material として添付されている統計解析計画書の Appendix B に記載されている例を示す。なお，この計算例は，実際の臨床試験と異なりステージ 1 とステージ 2 で同じ症例数が登録されることを想定している($n_{1p} = n_{2p}$)ことに注意が必要である。この場合，$w_1^2 = w_2^2 = 1/2$であるため，ステージ 1 とステージ 2 の併合 p 値は以下の式で求めることができる。

$$C(p, q) = 1 - \Phi[0.701\Phi^{-1}(1-p) + 0.701\Phi^{-1}(1-q)]$$

ステージ 1 の試験薬の各用量・用法群①から④とプラセボ群との対比較の p 値をp_1, p_2, p_3, p_4とし，その値は例えば 0.08, 0.04, 0.03, 0.02 であったとする。この時中間解析の結果，用量・用法群④が選択されたとし，ステージ 2 のみのデータでプラセボ群と比較して算出された p 値が 0.01 だったとする。これらの帰無仮説に基づく p 値は Simes 法で算出する。例えばH_{1234}に対する p 値p_{1234}は，

$$p_{1234} = \min(4 \times 0.02, 4/2 \times 0.03, 4/3 \times 0.04, 4/4 \times 0.08) = 0.053$$

で与えられる。この帰無仮説H_{1234}下での，ステージ 2 を含めた統合 p 値は

$$C(p_{1234}, q) = 1 - \Phi[0.701\Phi^{-1}(1-0.053) + 0.701\Phi^{-1}(1-0.01)] = 0.0027$$

となる。これを前述した残り 7 個の積仮説にも適用し，最も大きい p 値が有意水準 0.005 より小さい場合には帰無仮説が棄却されることとなる。

試験結果

中間解析時点は 188 例の患者が解析対象集団に含まれた。解析結果は以下の通りであった。安全性にも問題がなかったため，前述した B が勧告され 3 mg/kg/日・6 か月群が選択された。ただしデータモニタリング委員会の勧告が出るまでにプラセボ群，及び 4 つの試験薬群に対する症例登録は継続されたことから，結果としてデータモニタリング委員会の勧告時点でステージ 2 の目標症例数に達していたため，その時点で症例登録は終了された。

表 3-2-2-1 中間解析時の 6 か月後の有効率 (ITT 集団)

	プラセボ群 (25 例)	1 mg 3M 群 (41 例)	1 mg 6M 群 (40 例)	3 mg 3M 群 (39 例)	3 mg 6M 群 (43 例)
有効率(例数)	8.0% (2 例)	9.8% (4 例)	37.5% (15 例)	7.7% (3 例)	62.8% (27 例)

M: か月

審査報告書[18]をもとに作成

最終の解析結果は以下に示す通りであり，プラセボ群の有効率は 3.6%，3 mg/kg/日・6 か月群の有効率は 60.4%であり，有意な差が認められている。

表 3-2-2-2 最終解析時の 6 か月後の有効率 (ITT 集団)

	プラセボ群 (55 例)	3 mg 6M 群 (101 例)
有効率(例数)	3.6% (2 例)	60.4% (61 例)
調整 p 値(片側)*	< 0.0001	

M: か月

*: Posh et al.の統合検定法に基づく[19]

審査報告書[18]をもとに作成

有効性に関する審査上の論点

本試験結果に対して，PMDA では調査した範囲で有効性に関して大きな論点はなく，有効性が支持されている。一方，FDA では前述の通りデータモニタリング委員会で結論が出るタイミングでほとんど症例集積は終わっていたことから，アダプティブデザインではなく，5 群の並行群間比較試験として解釈できることが指摘されている[21]。その結果に基づくと，最終解析の有効率の点推定値は 1 mg と 3 mg で大きな差異がなかったことから，3 mg を選択した理由を問われた。スポンサー側はステージと用量群で交互作用が検出されなかったことから，中間解析前後で併合した結果を治療効果の推定値として解釈できることを説明した。すなわち，中間解析前後を併合した解析結果で，有効率の点推定値が 3 mg/kg/日群の方が高かったため，当該用量を採用したことの妥当性を主張したと推察される。

単群で実施された国内第 III 相試験（M703101-01）

M703101-01 試験は国内で実施された多施設共同単群試験である。目標症例数 30 例であり，1 mg/kg/日から投与を開始し，2 日ごとに 1 mg/kg ずつ（投与開始 3 日目に 2 mg/kg/日，投与開始 5 日目に 3 mg/kg/日）増量する試験デザインであった。なお，増量後の最終的な用量は，V00400 SB 2 01 で選択された用量と同じである。主要評価項目は 6 か月後における乳児血管腫に対する有効率であった。有効性判定の閾値は 12%と設定され，有効率の正確な 95%信頼区間下限が 12%を上回った時に成功と考えられた。なお閾値である 12%は前述した海外第 II/III 相試験（V00400 SB 2 01）で得られたプラセボ群の有効率から，最も高い場合を想定して算出されており[22]，海外試験におけるプラセボ群の有効率の 95%信頼区間の上限から算出されているのではないかと推測される。

M703101-01 試験の目標症例数は 30 例であったが，実際には 32 例が有効性解析対象集団に組み入れられた。すべての症例が 3 mg/kg/日で維持され，主要解析の結果，6 か月後の有効率［95%信頼区間］は 78.1 %（25/32 例）［60.0 %, 90.7 %］であり有効性が示されている。

本試験結果に対して，PMDA では調査した範囲で有効性に関して大きな論点はなく，有効性が支持されている。

まとめ

3.2.2 節では希少疾病で，アダプティブシームレス第 II/III 相デザインを用い，中間解析での用量選択，及びその有効性の検証を実施した事例を紹介した。希少疾病では多くの患者の集積が難しく，アダプティブデザインを用いることで，より効率的な試験を実施できる可能性がある。実際に試験を実施する場合には，どのような状況下でアダプティブデザインが効果的かを十分に吟味したうえで，規制当局との事前協議が必要になると考える。また，中間データで意思決定を行うことに起因するデータの完全性などの試験実施上の課題も存在するため，そのような統計的な観点以外の問題への配慮も必要であろう。

参考文献

[1] Gallo, Paul, et al. "Adaptive designs in clinical drug development—an executive summary of the PhRMA working group." Journal of biopharmaceutical statistics 16.3 (2006): 275-283.

[2] Food and Drug Administration. "Adaptive designs for clinical trials of drugs and biologics-Guidance for Industry." (2019).

[3] 日本製薬工業協会. "アダプティブデザインに関する FDA ガイダンスの邦訳." (2021) https://www.jpma.or.jp/information/evaluation/results/allotment/adaptive_design.html

[4] Lesaffre, Emmanuel, Gianluca Baio, and Bruno Boulanger, eds. Bayesian Methods in Pharmaceutical Research. CRC Press, 2020.

[5] Committee for Medicinal Products for Human Use. "Reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan." London: European Medicines Agency (EMEA) (2006).

[6] Jennison, Christopher, and Bruce W. Turnbull. Group sequential methods with applications to clinical trials. CRC Press, 1999.

[7] 森川敏彦・山中竹春 訳. "臨床試験における群逐次法：理論と応用." CAC, 2012

[8] Lawrence Gould, A., and Weichung Joseph Shih. "Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance." Communications in Statistics-Theory and Methods 21.10 (1992): 2833-2853.

[9] Gould, A. Lawrence. "Interim analyses for monitoring clinical trials that do not materially affect the type I error rate." Statistics in medicine 11.1 (1992): 55-66.

[10] Gould, A. Lawrence. "Planning and revising the sample size for a trial." Statistics in medicine 14.9 (1995): 1039-1051.

[11] Shih, Weichung Joseph. "Sample size re‐estimation‐journey for a decade." Statistics in Medicine 20.4 (2001): 515-518.

[12] Chow, Shein-Chung, and Mark Chang. Adaptive design methods in clinical trials 2nd edition.

Chapman and Hall/CRC, 2012.

[13] Bartlett, Robert H., et al. "Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study." Pediatrics 76.4 (1985): 479-487.

[14] Wei, L. J., and S. Durham. "The randomized play-the-winner rule in medical trials." Journal of the American Statistical Association 73.364 (1978): 840-843.

[15] Ware, James H. "Investigating therapies of potentially great benefit: ECMO." Statistical Science 4.4 (1989): 298-306.

[16] 平川 晃弘・五所 正彦(監訳) 臨床試験のためのアダプティブデザイン. 朝倉書店, 2018

[17] Berry, Scott M., et al. Bayesian adaptive methods for clinical trials. CRC press, 2010.

[18] 医薬品医療機器総合機構. "ヘマンジオルシロップ小児用, 審査報告書." (2016) http://www.pmda.go.jp/drugs/2016/P20160728001/730155000_22800AMX00431000_A100_2.pdf

[19] Posch, Martin, et al. "Testing and estimation in flexible group sequential designs with adaptive treatment selection." Statistics in medicine 24.24 (2005): 3697-3714.

[20] Léauté-Labrèze, Christine, et al. "A randomized, controlled trial of oral propranolol in infantile hemangioma." N Engl J Med 372 (2015): 735-746.

[21] Food and Drug Administration. "Propranolol. Statistical Review" (2014). https://www.accessdata.fda.gov/drugsatfda_docs/nda/2014/205410Orig1s000StatR.pdf

[22] Kaneko, Tsuyoshi, et al. "Efficacy and safety of oral propranolol for infantile hemangioma in Japan." Pediatrics International 59.8 (2017): 869-877.

3.3 エンリッチメント戦略

3.3.1 概要

FDA の Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products Guidance for Industry[1]ではエンリッチメントとは前向きに，より治療効果を示しやすい患者の集団を選択すること，とされており大きく 3 つに大別される。

1. 異質性を減らすための戦略: ベースラインで疾患に関する観測値または疾患を特徴づけるバイオマーカーの観測値が一定の範囲にある患者の選択（患者間変動の減少），及び疾患・症状が自然に改善する患者や測定値の変動が大きい患者の除外（患者内変動の減少）などが考えられる。これらの方策によって得られるばらつきの減少は，試験の検出力を高めると考えられる。
2. 予後的エンリッチメント戦略: 疾患関連のイベントが発現する可能性が高い患者（イベントを評価する試験の場合）または状態が悪化する可能性が高い患者（評価項目が連続量である場合）の選択。これらの戦略は群間の絶対効果差を増加させるが，相対効果を変化させることはないと考えられる。例えばある疾患において，全患者を対象にして試験を実施した場合に，治療 A と治療 B の死亡率がそれぞれ 5%と 10%であるのに対して，特定の予後因子を持つ患者で臨床試験をした場合に死亡率が 20%と 40%である場合などが考えられる。主要評価項目が二値応答の試験では群間差が大きくなり，症例数を減らすことができる。また，主要評価項目がハザード比のような Time to event 型の試験でも，よりイベントが発生する患者を対象に試験を実施することで，症例数を減らすことができる。
3. 予測的エンリッチメント戦略: 治療中の疾患を有する他の患者よりも薬物治療に反応する可能性が高い患者を選択することが含まれる。このための手法として①経験的なストラテジー，②病態生理学的ストラテジー，③経験的なゲノムストラテジー[2]，④ランダム化中止試験 [3]，⑤ノンレスポンダーや他の治療に忍容性がない患者を対象にした試験，が挙げられている。

これらのエンリッチメント戦略に関しては，FDA のガイダンスにもいくつか具体例が記載されているが，予後的エンリッチメント戦略，及び予測的エンリッチメント戦略に関しては FDA の Rare Diseases: Common Issues in Drug Development Guidance for Industry[2]でも言及されている。また希少疾病での適応に関してはライソゾーム病での適応事例が Boudes et al. (2013) [3]を始め様々な報告がされている。報告されている事例を表 3-3-1-1 にまとめた。異質性を減らすための戦略，予後的エンリッチメント戦略に関しては明確に「エンリッチメント」という単語を使わずに，臨床試験における適格・除外基準で設定されているケースもあり，実際には他にも適応事例は存在すると考

2 予後と関連すると思われる特定のゲノムパターン（例:RNA 発現プロファイルなど）を示す患者サブセットを対象とした試験。治療効果の差に対する病態生理学的根拠を提供することなく，より治療に反応性を示すサブセットを特定することができる。
3 ランダム化中止試験では，非盲検期間中の治療に反応を示した患者が薬物治療の継続又はプラセボ治療のいずれかにランダム化される。治療に反応したと思われる患者のみを対象とするため，エンリッチした試験と考えることができる。

えられる。

以下では予測的エンリッチメント戦略を用いた臨床試験として，Mehta et al. (2019) [4]で報告されている進行性血管肉腫を対象にアダプティブエンリッチメントデザインが FDA と EMA の議論に従い計画された（FDA では special protocol assessment も実施），carotuximab の TAPPAS 試験デザインについて紹介する。なお，後述の通り本試験は中間解析の時点で中止され，承認申請には至っていない。

表 3-3-1-1 希少疾病でエンリッチメント戦略が用いられた，もしくはガイダンスで提案されている事例

エンリッチメント戦略の種類	疾患名	内容
異質性を減らすための戦略	ムコ多糖症 I 型	6 分間歩行試験で，正常または正常に近い患者は改善が認められる可能性が低く，また重度な患者は反応が限定的であるため除外している[3]
予後的エンリッチメント戦略	ポンペ病	疾患進行が長期になる成人ではなく，余命が短い重症な小児を対象に試験を実施する場合がある[3]
	肺動脈性肺高血圧症	いくつかの POC 試験で COMPERA, French スコア, REVEAL リスクスコアがより悪化に関するイベントを起こす可能性の高い患者を特定できる可能性を示している。肺動脈性肺高血圧症の病態，及び病態生理に関する現在の理解から，患者のベースライン時の死亡率に関する個々のリスクにかかわらず治療効果を支持しているため，FDA はこの疾患領域で治療効果を低リスク群に外挿することに難色を示さないことをコメントしている[5]
	デュシェンヌ型筋ジストロフィー	疾患進行を遅くするが，改善や既に起こった筋肉の機能不全を戻すことが期待されていない薬剤では，登録基準でより疾患進行が速いと予測される患者を組み入れることも考えられる。ただしエンリッチメント戦略を行うために，科学的な妥当性なしに患者を不必要に除外することはするべきでない[6]
	常染色体優性多発性嚢胞腎	両腎容積を年齢とベースライン eGFR と併せて用いることで，腎機能低下のハイリスク患者を特定可能[7]
	進行性線維化を伴う間質性肺疾患	UIP 様の PF-ILD を持つ患者は他の繊維化パターンを持つ患者より悪化率が高く，治療効果が大きくなると想定[8]

エンリッチメント戦略の種類	疾患名	内容
予測的エンリッチメント戦略	嚢胞性線維症	Ivacaftor は，薬物に反応を示すことができる CFTR 遺伝子の p-G551D 変異がある患者のみに承認されている[3]
	ファブリー病	ミガーラスタット塩酸塩（ガラフォルド®）はより反応性が高いであろう GLA 遺伝子変異を伴う患者を対象に臨床試験が実施され[3]，GLA 遺伝子変異を伴う患者に対して承認がされている。
	慢性炎症性脱髄性多発神経炎	ランダム化中止試験を実施 [9]
	進行性血管肉腫	アダプティブエンリッチメントデザインを用いた臨床試験実施[4]

3.3.2 Carotuximab

進行性血管肉腫について

進行性血管肉腫は米国では年間約 1,000 例，欧州でも同程度の発生率である超希少腫瘍である。切除不能進行性血管肉腫患者に対する標準治療には，化学療法（タキサン系薬剤，アントラサイクリン系薬剤，及びゲムシタビン）のほか，血管内皮増殖因子受容体（VEGFR）を含む複数の受容体チロシンキナーゼ阻害剤であるパゾパニブが含まれる。これらの治療法による転移性疾患の腫瘍コントロールは短期間であり，無増悪生存期間（PFS）の中央値は 3.0~6.6 か月，全生存期間（OS）の中央値は約 8~11 か月である。転移性疾患である進行性血管肉腫における無増悪期間は短く，より効果的で忍容性の高い治療選択肢が必要であった。

Carotuximab（TRC105）について

Carotuximab（別名 TRC105）はエンドグリン（CD 105）に対するモノクローナル抗体である。Carotuximab は先行して実施された第 I/II 相試験でパゾパニブとの併用で持続的な完全奏効が観測され，無増悪生存期間の中央値は 7.8 か月間であった。これはパゾパニブ単剤で 40 例に対して実施された臨床試験の結果（完全奏効例なし，無増悪生存期間の中央値 3.0 か月，全生存期間の中央値 9.9 か月）と比較して良好な結果であった。また皮膚病変のある患者で Carotuximab のより高い効果が示唆されていた。

アダプティブエンリッチメントデザインを用いた TAPPAS 試験

上記の結果を受けて第 III 相試験である TAPPAS 試験が計画された。TAPPAS 試験はランダム化国際共同多施設非盲検並行群間試験である。EMA と FDA との議論に従い計画された試験であり，FDA では special protocol assessment も実施されていた。被験者は Carotuximab とパゾパニブの併用療法群とパゾパニブ単剤群に 1:1 の割付比で割付けられた。割付にあたり層別因子として①皮膚病変の有無，②前治療ライン数（0 vs 1 か 2）が設定された。主要評価項目は無増悪生存期間であった。主要解析方法に関しては明記されていないが，症例数設定にハザード比を用いていることから，log rank 検定か Cox 比例ハザードモデルであったと推察される。

中間解析

進行性血管肉腫は非常に稀な疾患であること，及び事前に得られている情報が限られていることから，非盲検下での中間解析に基づくアダプティブデザインが計画された。試験中のデザイン変更としては，適切な検出力を持つように症例数を増加させる症例数再設定，及び中間解析の結果によっては皮膚病変のある患者のみにその後の登録を制限することを可能にする予測的エンリッチメント戦略が計画された。具体的には中間解析時の条件付き検出力に基づき，データモニタリング委員会は以下の 4 つの Zone のいずれかに基づく症例の継続的な登録を推奨する。このデザインは Mehta and Pocock (2011) [10]の Promising zone に基づく臨床試験に，Enrichment zone を追加する形でエンリッチメント戦略を追加した試験と考えることができる。

A. Favorable zone: もともとの計画のまま試験を続行

B. Promising zone: 症例数，及び最終解析時の PFS イベント数を増やして試験続行

C. Enrichment zone: 皮膚病変のある患者のみ登録を続行

D. Unfavorable zone: もともとの計画のまま試験を続行

今回 Time to event 型である無増悪生存期間(PFS)を用いているため，エンリッチメント戦略をとる際の留意点が生じる。試験中のデザイン変更の統計的妥当性は，中間解析に用いるデータセットが試験中のデザイン変更によってもたらされる結果と独立であることが必要である。Time to event 型のエンドポイントを主要評価項目とした臨床試験では，中間解析時にイベントが起こっておらず，中間解析時に打ち切りと扱われた患者が後々イベントを起こす場合がある。もしもエンリッチメントに関する決定（皮膚病変のある患者のみの登録へのデザイン変更）が，イベントと関連があるかもしれない患者のプロファイル（毒性，腫瘍縮小の程度）も考慮した上で決定された場合，第 1 種の過誤確率が増大する可能性がある。これは，例えば中間解析時にイベントが起きていない皮膚病変患者において，Carotuximab + パゾパニブ併用群でパゾパニブ単剤群と比較して腫瘍が縮小していた場合に，皮膚病変がある患者では今後 PFS のイベントが単剤群でより起こるであろうことが予測され，そういったデータに基づきエンリッチメント戦略へのデザイン変更が行われるとバイアスが生じる，という状況を想定していると考えられる。実際に，データモニタリング委員会は PFS のイベントだけでなく，その時点で利用可能なすべてのデータに基づき勧告を出す。こういったジレンマを解消するために Jenkins が提案したアプローチ[11]が用いられた。このアプローチでは中間解析で盲検解除される前に決められた症例数・イベント数に基づき試験データを 2 つのコホート（コホート 1，及びコホート 2）に分割する。中間解析時点でコホート 1 ではまだイベント数を集積中である。データモニタリング委員会の症例数・イベント数に関する勧告はコホート 2 にのみ適応され，コホート 1 は事前に規定したイベント数に達するまで，デザインの変更なく試験を続行する。これにより統計的に妥当な p 値を得ることができる。以上を踏まえ，最終的な試験デザインは試験の動作特性も考慮して以下のように定められた。

コホート 1

120 例の症例が登録される。中間解析はコホート 1 で 40 例のイベントが観測された時点または 120 例目の症例で登録から 30 日経過した時点で実施される。中間解析時に，全体集団での条件付き検出力 CP_F と皮膚病変がある患者での条件付き検出力 CP_S をそれぞれ算出する。中間解析でのデータモニタリング委員会の勧告に依らず，コホート 1 で 60 例イベントが発生するまで追跡を行う。

コホート 2

・$CP_F > 95\%$の場合（Favorable zone）：デザインに変更はなく，コホート 2 として全体集団で 70 例の患者を登録し，コホート 2 で 35 イベントが発生するまで追跡を行う。最終解析は全体集団を対象に検定を行う。

・30% ≦ CP_F ≦ 95%の場合（Promising zone）: コホート 2 における症例数を 220 例に増加させ，コホート 2 で 110 イベントが発生するまで追跡を行う。最終解析は全体集団を対象に検定を行う。
・CP_F < 30%, 及び CP_S ≧ 50%の場合（Enrichment zone）: コホート 2 で皮膚病変がある患者のみを 160 例登録し，皮膚病変がある患者においてコホート 1，及びコホート 2 の合計で 110 イベントが発生するまで追跡を行う。最終解析は皮膚病変がある患者集団を対象に検定を行う。
・CP_F < 30%, 及び CP_S < 50%の場合（Unfavorable zone）: デザインに変更はなく，コホート 2 として全体集団で 70 例の患者を登録し，コホート 2 で 35 イベントが発生するまで追跡を行う。最終解析は全体集団を対象に検定を行う。

最終解析での検定方法に関して述べる。中間解析で集団選択が実施されるため多重性の調整が必要である。まず全体集団での帰無仮説をH_0^F，皮膚病変がある患者での帰無仮説をH_0^S，これらの積仮説を$H_0^{FS} = H_0^F \cap H_0^S$とする。

またコホート 1 での全体集団，皮膚病変がある患者集団での片側 p 値をそれぞれp_1^F，及びp_1^S，コホート 2 での全体集団，及び皮膚病変がある患者集団での片側 p 値をそれぞれp_2^F，及びp_2^Sとする。この時コホート 1，及びコホート 2 での積仮説に対する p 値は Simes 法を用いてそれぞれ$p_1^{FS} = \min[2\min\{p_1^F, p_1^S\}, \max\{p_1^F, p_1^S\}]$，及び$p_2^{FS} = \min[2\min\{p_2^F, p_2^S\}, \max\{p_2^F, p_2^S\}]$で与えられる。この時中間解析でエンリッチメントがされなかった場合は全体集団に対して最終解析で検定を行うので，閉検定手順により

$$p^{FS} = 1 - \Phi\{w_1\Phi^{-1}(1 - p_1^{FS}) + w_1\Phi^{-1}(1 - p_2^{FS})\}$$
$$p^F = 1 - \Phi\{\mathrm{w}_1\Phi^{-1}(1 - p_1^F) + \mathrm{w}_1\Phi^{-1}(1 - p_2^F)\}$$

で得られるp^{FS}，及びp^Fがいずれも片側有意水準の 0.025 を下回る必要がある（なお原著ではp_2^{FS}であるがp_2^Fではないかと思われる)。ここで$w_1 = \sqrt{d_1/(d_1 + d_2)}$，$w_2 = \sqrt{d_2/(d_1 + d_2)}$であり，$d_1$，及び$d_2$はコホート 1，及びコホート 2 で計画されているイベント数である（$d_1 = 60$，$d_2 = 35$)。これは症例数再設定が行われても同じである。一方，中間解析でエンリッチメントした場合は同じく閉検定手順により

$$p^{FS} = 1 - \Phi\{w_1\Phi^{-1}(1 - p_1^{FS}) + w_1\Phi^{-1}(1 - p_2^S)\}$$

$p^S = 1 - \Phi\{w_1\Phi^{-1}(1 - p_1^S) + w_1\Phi^{-1}(1 - p_2^S)\}$ で得られるp^{FS}，及びp^Sがいずれも片側有意水準の 0.025 を下回る必要があった。

このデザインに基づいたシミュレーションの結果は Mehta et al. (2019) [4]の Table 1 に記載されている。ハザード比が全体集団と皮膚病変がある患者で同様である場合には，アダプティブデザインよりも固定デザインで検出力が高かったが，2 つの集団でハザード比の乖離がある場合はアダプティブデザインの方で検出力が高くなった。これは集団間の異質性が限定的であると想定できる状況では固定デザインを用いたほうが効率的であることを示しているが，潜在的な異質性が想定できる状況ではアダプティブデザインを用いたほうが検出力の観点から好ましい場合があることを示している。論文中には第 1 種の過誤確率に関するシミュレーション結果の記載はないが，閉検定手順を用いることで理論的に第 1 種の過誤確率も制御されていると考える。

試験結果

中間解析の結果，コホート1終了時点と思われる123例で最終解析を実施することが決定され，主要評価項目であるPFSの中央値がCarotuximab+パゾパニブ併用群で4.2か月，パゾパニブ単剤群で4.3か月であり，ハザード比，及びその信頼区間は0.98 [0.52 – 1.8]と，PFSの延長が見られなかったことが報告されている[12]。

まとめ

3.3.2節では希少疾病において，エンリッチメント戦略が用いられた事例について紹介した。医薬品開発に本戦略をとることで，多くの症例を集積することが難しい希少疾病において効率的な試験実施が可能になると考えられる。一方，エンリッチメント戦略を用いた場合には，患者集団の一部のみを対象に臨床試験を実施することが問題点となる場合がある。例えばランダム化中止試験を実施した場合，治療に対して反応があった患者のみを対象に比較試験を実施する。しかし実際の臨床現場では，患者が治療に反応性はあるかないかは投与するまでわからない。そのためこれらの戦略をとる際には，試験結果の一般化可能性についても十分に考慮する必要があるであろう。また，患者背景によって治療効果の違いがある可能性が示唆されている場合には，Carotuximabの事例のようにアダプティブエンリッチメントデザインを適用した試験の実施も考えられる。

参考文献

[1] Food and Drug Administration. "Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products - Guidance for Industry." (2019).

[2] Food and Drug Administration. "Rare Diseases: Common Issues in Drug Development - Guidance for Industry (Draft)." (2019).

[3] Boudes, Pol F. "Clinical studies in lysosomal storage diseases: Past, present, and future." Rare Diseases 1.1 (2013): e26690.

[4] Mehta, C. R., L. Liu, and C. Theuer. "An adaptive population enrichment phase III trial of TRC105 and pazopanib versus pazopanib alone in patients with advanced angiosarcoma (TAPPAS trial)." Annals of Oncology 30.1 (2019): 103-108.

[5] Garnett, Christine, and Norman Stockbridge. "Ask the expert: a regulatory perspective on clinical trials for pulmonary arterial hypertension." Advances in Pulmonary Hypertension 19.2 (2020): 62-65.

[6] Food and Drug Administration. "Duchenne Muscular Dystrophy and Related Dystrophinopathies: Developing Drugs for Treatment - Guidance for Industry." (2018).

[7] Food and Drug Administration. " Qualification of Biomarker Total Kidney Volume in Studies for Treatment of Autosomal Dominant Polycystic Kidney Disease - Guidance for Industry (Draft)." (2016).

[8] Flaherty, Kevin R., et al. "Design of the PF-ILD trial: a double-blind, randomised, placebo-controlled phase III trial of nintedanib in patients with progressive fibrosing interstitial lung disease." BMJ open respiratory research 4.1 (2017): e000212.

[9] van Schaik, Ivo N., et al. "Subcutaneous immunoglobulin for maintenance treatment in chronic inflammatory demyelinating polyneuropathy (PATH): a randomised, double-blind, placebo-controlled, phase 3 trial." The Lancet Neurology 17.1 (2018): 35-46.

[10] Mehta, Cyrus R., and Stuart J. Pocock. "Adaptive increase in sample size when interim results are promising: a practical guide with examples." Statistics in medicine 30.28 (2011): 3267-3284.

[11] Jenkins, Martin, Andrew Stone, and Christopher Jennison. "An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints." Pharmaceutical statistics 10.4 (2011): 347-356.

[12] Jones, R. L., et al. "Results of the TAPPAS trial: An adaptive enrichment phase III trial of TRC105 and pazopanib (P) versus pazopanib alone in patients with advanced angiosarcoma (AS)." Annals of Oncology 30 (2019): v683.

3.4 マスタープロトコル

3.4.1 概要

マスタープロトコルは，FDA のガイダンスにおいて，”a master protocol is defined as a protocol designed with multiple substudies, which may have different objectives and involve coordinated efforts to evaluate one or more investigational drugs in one or more disease subtypes within the overall trial structure. “と定義されており，例えば，複数の治療法について同時対照群を共有する，一部の治療方法については途中のデータに基づいて脱落させる，というように，事前に規定したフレームワークの下で複数の薬剤・複数の疾患の評価を行うことを可能にすると考えられる[1]。本質的な試験特性を特徴づけるための試験デザインとして，以下のような目的を持つアンブレラ，バスケット，プラットフォームと呼ばれる試験がある。

- アンブレラ試験：単一とみなせる対象疾患の中で，複数の治療方法について評価することを意図した試験デザイン（例：図 3-4-1-1）
- バスケット試験：興味のあるひとつの治療方法を複数の疾患または疾患のサブタイプごとに評価することを意図した試験デザイン（例：図 3-4-1-2）
- プラットフォーム試験：意思決定アルゴリズムに従った治療方法の新規追加・除外を計画段階から許容しつつ，単一とみなせる対象疾患の中で，複数の治療方法について永続的に評価することを意図した試験デザイン（例：図 3-4-1-3）

図 3-4-1-1，3-4-1-2，3-4-1-3 は，いずれも，Woodcock et. al. (2017)からの改変である[2]。なお，マスタープロトコルに基づく試験は上述の 3 つの試験に完全に区分できるとは限らず，その複数の特徴を有するという場合もあるだろう。マスタープロトコルに基づく試験としては，Lung-MAP 試験[2, 3]と ISPY-2 試験[2, 4, 5]を補遺にて簡単に紹介するが，例えば，Lung-MAP 試験は，アンブレラ試験と同時にプラットフォーム試験の特徴も有しているとも考えられる。

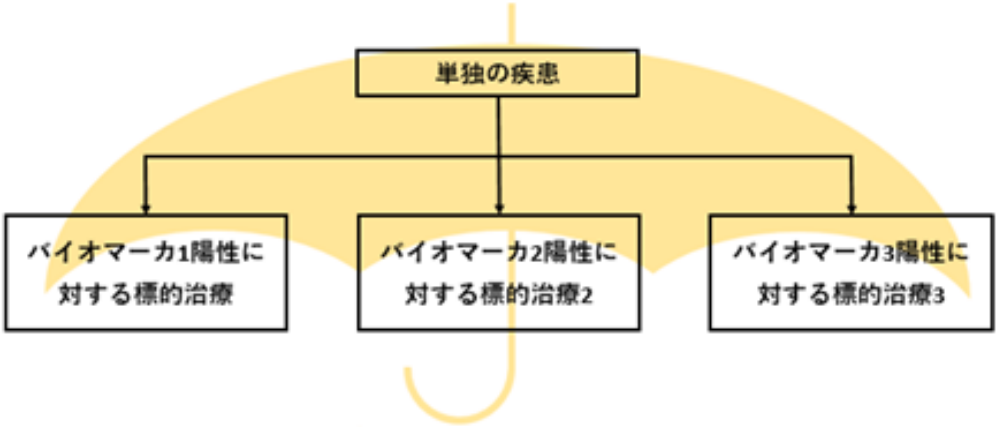


図 3-4-1-1　アンブレラ試験のイメージ図

この図の例では，ある疾患を有する患者に対して，特定のバイオマーカーを検査し，結果によって対応するサブスタディに割り付ける。個々のサブスタディ毎で異なる治療法を評価する。例えば，進行期非小細胞肺がん（扁平上皮癌）の患者を対象とする Lung-MAP 試験はこの特徴を有すると考えられる[3]。

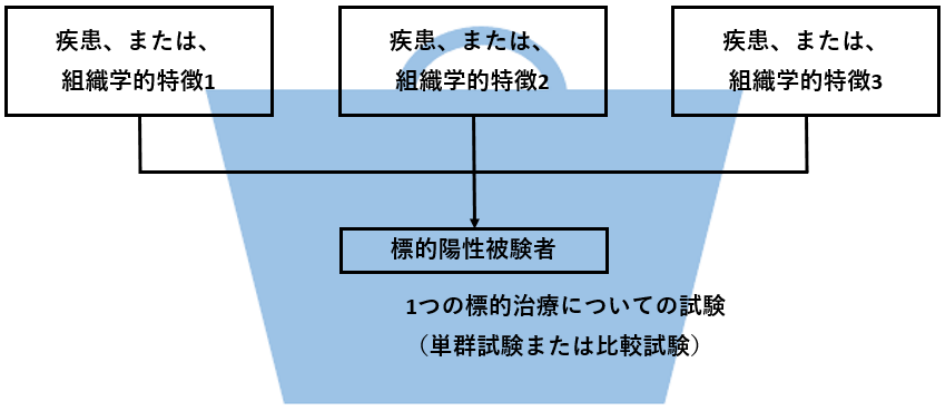


図 3-4-1-2　バスケット試験のイメージ図

この図の例では，疾患や組織学的特徴（例えば，がん種）を有する患者に対して特定のバイオマーカーを検査し，結果によって試験に組み入れる。メラノーマ以外で BRAFV600 変異を有する 9 つのがん種に対して，ベムラフェニブの有効性を評価した BRAF-V600 試験はこの特徴を有すると考えられる[6]。

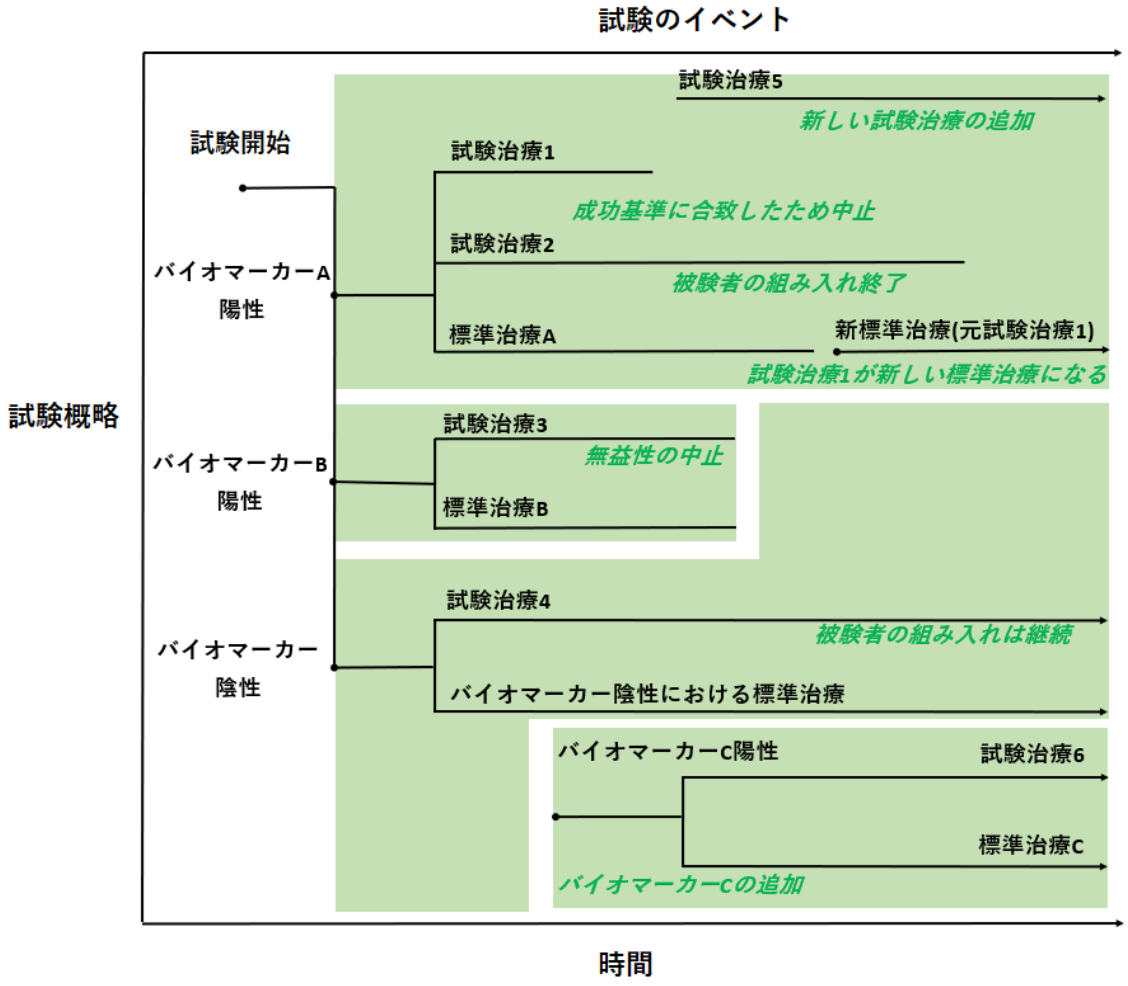


図 3-4-1-3　プラットフォーム試験のイメージ図

この図の例では，横軸を時間軸とし，試験治療の追加・脱落や新たなサブスタディの追加などの変遷を示した。この試験の開始した時では，スクリーニング時にバイオマーカーA，B を検査し，組入れられた患者はそのスクリーニング検査の結果によって，「バイオマーカー A 陽性」「バイオマーカーB 陽性」「バイオマーカー陰性」の 3 つのグループのうち 1 つに割り当てられる。「バイオマーカーA 陽性グループ」では，試験治療 1，2 と標準治療 A のいずれかが割り付けられる。試験が進行するうちに，新しい治療が利用可能となれば，新しい試験治療として追加される（この図では，試験治療 5 が新たに加わり，試験治療 1，2，5 を標準治療 A と比較する試験に途中から変更されている）。また，試験が進行するうちに，評価していた試験治療が新しい標準治療となるかもしれない（この図では標準治療 A が途中から新しい標準治療として試験治療 1 に置き換わっている）。「バイオマーカーB 陽性グループ」では，試験治療 3 が標準治療 B に対してのベネフィットが示せずに，このグループでの評価は中止となり，この時点でバイオマーカーB 陽性の患者は「バイオマーカー陰性グループ」に組み込まれる。試験が進行し，バイオマーカーC や評価対象となる治療（試験治療 6，標準治療 C）がこの試験でも利用可能になった時点で「バイオマーカーC 陽性グルー

プ」が新たに追加される。即ち，この時点では，スクリーニング時にバイオマーカーA, C を検査し，組入れられた患者はそのスクリーニング検査の結果によって，「バイオマーカーA 陽性」「バイオマーカーC 陽性」「バイオマーカー陰性」の 3 つのグループのうち 1 つに割り当てられる。ここで記載した試験デザインはあくまで 1 つの仮想的な例であるが，例えば ISPY-2 試験のように，レスポンスアダプティブランダム化（3.2 節）を取り入れることも考えられる。

マスタープロトコルに基づく臨床試験が注目されている理由について，Woodcock et. al. (2017)は単一の対象疾患に 1 つの試験目的を設定した旧来の臨床試験では費用や実施可能性の面で限界があり，患者集団に対してより多くのクリニカルクエスチョンを，より効率的かつ短期間で得ることが近年は求められていることを指摘している。同論文では，マスタープロトコルの革新性について，（1）インフラストラクチャと（2）試験デザインの点から考察しているが，いずれの点からも比較的患者数の多い疾患だけでなく希少疾患においても有効な試験デザインとして期待できると指摘している[2]。

- インフラストラクチャ

共通のプロトコルを用いることで，個別で試験を実施していた場合よりも患者が試験参加の機会を損なう可能性が低くなると考えられる。例えば，個別の試験では参加しようとした試験の組み入れ基準に合致しない可能性もあるが，マスタープロトコルに基づく試験であれば，共通のスクリーニングプラットフォーム下で患者が参加できるサブスタディを特定できるかもしれない。また，治験審査委員会やデータモニタリング委員会などの各種委員会を一元化できるといった試験のガバナンスの観点からもマスタープロトコルを活用するメリットが考えられる。この他にも，同じプラットフォームの中でより効率的に均質なデータを収集できることも，マスタープロトコルに基づく試験のメリットと考えられる。例えば，平川ら（2019）も，がん領域において，がん種や進行度，組織型等により定義した部分集団ごとに従来の枠組みでの臨床試験を実施すると作業重複が多く効率的ではない場合も考えられること，共通のプラットフォームを用いることによりこういった非効率性の解消が期待される点を指摘している[8]。

- 試験デザイン

まず，デザインや治験実施計画書の手技などについて個々の治療方法以外（例えば，来院時期や測定の方法，アウトカムの定義など）は基本的に同じになるというメリットがある。デザイン自体もアダプティブデザインなどの革新的な方法を反映することで新規治療の追加・脱落・次相への移行といった試験デザインの変更を可能とし，あるいは，対照群の共有，得られたデータを自然歴研究として将来的な活用，複数のリサーチクエスチョンを効率的に満たすといったことが期待される。ただし，これらの柔軟性は，マスタープロトコルにつ

いての事前の適切な計画が大前提であり，事前に計画されていないデザイン変更によるものではない。

上記の革新性によって，試験の計画と運用・管理が一元化され，被験者の効率的な登録，開発の迅速化，また，被験者にとっても試験に参加する機会が高まるという点が期待される一方で，マスタープロトコルに基づく試験の実施に当たってはいくつかの留意点がある。

まずインフラストラクチャを整えるための資源，事前の綿密な計画と様々な関係者・委員会等の調整と合意形成，運営やガバナンスなど，単独の試験と比べるとより時間を要する点である。より複雑なデザインや意思決定を予定する場合は，より多くの事前検討が必要となろう。平川ら (2019) はアンブレラ試験やプラットフォーム試験では，大規模かつ長期間のマスタープロトコル試験になるため，試験の管理・運用面の負担が大きく，永続的な試験運用を行う組織，及び枠組みの構築が課題となる点も指摘している。解析手法に関してもリサーチクエスチョンと合致したものであるべきであり，誤った結論を導くことのないよう検討を続ける必要がある。複数のサブ試験を取り扱うことから，多重性の問題や意思決定基準など，シミュレーションに基づく詳細な検討が必要となることが考えられる[8]。

もう一つの留意点は，比較的実施期間の長いマスタープロトコルでは，その期間中に対象疾患の標準治療が変更されることによってデザインも影響される可能性がある点である。例えば，Lung-MAP 試験では，その実施期間中にニボルマブがドセタキセルに対する優越性を示し，また，ISPY-2 試験でも，その実施期間中に試験治療のひとつであったペルツズマブとトラスツズマブと化学用法の併用療法がトラスツズマブと化学療法の併用療法に代わって HER2 陽性乳がんの標準治療となった[2]。これらの市場の変化に応じて ISPY-2 試験，及び Lung-MAP 試験はマスタープロトコルを更新しデザインを変更している。変更に当たっては，対照群の変更に起因する解析計画の変更や，より効果の高い標準治療と比較することについて検討するため，試験を一時的に止める必要性も考えられる[2]。

マスタープロトコルに基づく試験については日本でも計画・実施が進みつつある。2021 年に発出された「抗悪性腫瘍薬の臨床評価方法に関するガイドライン」では，マスタープロトコルは希少がん，希少なサブタイプに対する抗悪性腫瘍薬の開発を促進できるものと期待されている[7]。また，国立がんセンターが開始した希少がんに関する産官学の共同開発基盤である MASTER Key Project では，2017 年 5 月から 2022 年 8 月時点までで延べ 20 程度のサブ試験が症例登録を開始，または試験を完了している[9, 10]。次節では，日本における承認申請事例として，マスタープロトコル（バスケット試験）を用い，希少フラクションでかつがん種横断的に承認を取得した二つの薬剤について紹介する。また，承認申請に用いられた事例ではないものの，Washington University School of Medicine が実施する優性遺伝性アルツハイマーを対象としたプラットフォーム試験における solanezumab と gantenerumab の解析結果（NCT04623242）について，いくつか文献に基づき 3.4.4 節にて紹介する。

参考文献

[1] Food and Drug Administration. "Master Protocols: Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologics - Guidance for Industry." (2022).

[2] Woodcock, Janet, and Lisa M. LaVange. "Master protocols to study multiple therapies, multiple diseases, or both." New England Journal of Medicine 377.1 (2017): 62-70.

[3] Steuer, CE1, et al. "Innovative clinical trials: the LUNG‐MAP study." Clinical Pharmacology & Therapeutics 97.5 (2015): 488-491.

[4] Wang, Haiyun, and Douglas Yee. "I-SPY 2: a neoadjuvant adaptive clinical trial designed to improve outcomes in high-risk breast cancer." Current breast cancer reports 11.4 (2019): 303-310.

[5] Park, John W., et al. "Adaptive randomization of neratinib in early breast cancer." New England Journal of Medicine 375.1 (2016): 11-22.

[6] Hyman, David M., et al. "Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations." New England Journal of Medicine 373.8 (2015): 726-736.

[7] 厚生労働省. "抗悪性腫瘍薬の臨床評価方法に関するガイドライン." (2021) https://www.mhlw.go.jp/hourei/doc/tsuchi/T210401I0060.pdf

[8] 平川晃弘ら. "マスタープロトコルに基づくがん臨床試験." 計量生物学 39.2 (2019): 85-101.

[9] 国立研究開発法人国立がん研究センター MASTER KEY PROJECT（2022 年 9 月時点） https://www.ncc.go.jp/jp/ncch/masterkeyproject/index.html

[10] Okuma, Hitomi S., et al. "MASTER KEY project: powering clinical development for rare cancers through a platform trial." *Clinical Pharmacology & Therapeutics* 108.3 (2020): 596-605

3.4.2　エヌトレクチニブ（ロズリートレク®）

NTRK 融合遺伝子について

遺伝子の融合は発癌等における主要な原因の一つであり，慢性骨髄性白血病での Breakpoint cluster region-Abelson（*BCR-ABL*）融合遺伝子，非小細胞肺癌（NSCLC）における *ALK* 融合遺伝子，及び *ROS1* 融合遺伝子等が癌のドライバーとして働くことが報告されている[1, 2]。受容体型チロシンキナーゼである TRK をコードする *NTRK* 遺伝子が他の遺伝子と融合することにより，MAPK 経路等の下流シグナル伝達経路を恒常的に活性化する TRK 融合タンパクが産生される。当該融合タンパクは，発癌等における主要な原因の一つであり，腫瘍細胞の増殖・生存や正常細胞の腫瘍化に寄与していることが報告されており，患者数の多い結腸・直腸癌，NSCLC 等から患者数の少ない卵巣癌，軟部組織肉腫等，さらには小児癌である先天性線維肉腫等を含め，様々な癌で確認されている[3, 4]。なお，公表論文等で報告されている各癌種における *NTRK* 融合遺伝子の陽性率は一般に患者数の多い癌で低く，希少な癌で高い傾向にあるため，患者数は極めて少ないとされている。

エヌトレクチニブは TRK 等のチロシンキナーゼを阻害する低分子化合物であり，TRK（TRKA，TRKB，及び TRKC）等のリン酸化を阻害し，下流のシグナル伝達分子のリン酸化を阻害することにより，腫瘍増殖抑制作用を示すと考えられている。

エヌトレクチニブについて

海外において *NTRK*，*ALK* または *ROS1* 融合遺伝子陽性等の進行・再発の悪性固形腫瘍患者を対象とした第I相試験（ALKA 試験）が実施された。 前述のように，*NTRK* 融合遺伝子は特に患者数の少ないがんにおいて発現頻度が高い癌のドライバー遺伝子であり，癌腫ごとのランダム化比較試験の実施が困難であるため，*NTRK* 融合遺伝子陽性の進行固形がん患者（複数の癌腫）を対象としたバスケット試験として国際共同第II相試験（STARTRK-2 試験）が実施された。米国，及び欧州では，STARTRK-2 試験を主要な臨床成績として，それぞれ 2018 年 12 月，及び 2019 年 1 月に承認申請が行われた。米国では FDA によりブレークスルーセラピーに，欧州では EMA より PRIME（PRIority Medicines）に指定された（それぞれ 2019 年 5 月 15 日，同 10 月）。

なお，エヌトレクチニブは，2018 年 9 月に「前治療後に疾患が進行，または許容可能な標準治療がない NTRK 融合遺伝子陽性の局所進行または遠隔転移を有する成人，及び小児固形がん患者の治療」を予定される効能・効果として先駆け審査指定制度の対象品目に，また 2018 年 12 月に「*NTRK* 融合遺伝子陽性の局所進行または遠隔転移を有する固形がん」を予定される効能・効果として希少疾病用医薬品にそれぞれ指定され，2019 年 6 月 18 日に世界で初めて薬事承認を日本で取得している。以下ではエヌトレクチニブの審査報告書[5]に基づき，実施された臨床試験及びその結果に対する機構の見解について述べる。

エヌトレクチニブの臨床試験

エヌトレクチニブの有効性の評価に用いられた試験は国際共同第II相バスケット試験

（STARTRK-2 試験）であった。STARTRK-2 試験では *NTRK*, *ALK*, *ROS1* 融合遺伝子陽性の進行・再発の固形癌患者（18 歳以上）（目標症例数：①*NTRK* 融合遺伝子陽性の固形癌コホート 62 例, ②*ROS1* 融合遺伝子陽性の NSCLC コホート 150 例, 及び③その他のコホート 62 例）を対象に, エヌトレクチニブの有効性, 及び安全性を検討することを目的とした単群試験が, 日本を含む 15 の国で実施された。なお, 本試験では, 化学療法歴に係わらず症例登録が可能とされた。用法・用量は, エヌトレクチニブ 600 mg を QD で経口投与することとされ, 疾患進行または試験中止基準に該当するまで投与することとされた。

上記のうち, *NTRK* 融合遺伝子陽性の固形癌コホートに登録された 63 例のうち, 51 例が有効性の解析対象とされた。本試験の主要評価項目は, RECIST ver1.1 に基づく独立中央判定委員会（BICR）判定による奏効率とされた。また, 本試験では, 2 段階デザインが計画され, 閾値奏効率は 20%と設定された（試験計画時に *NTRK* 融合遺伝子について報告されていた癌種に対して二次治療以降等として使用可能な薬剤の奏効率が 0～41.6%(中央値：13.4%)の範囲で分布していたことを参考として設定された）。第一段階では最大 13 例の患者を対象に解析を実施し, うち, 奏効例が 3 例以下の場合に症例登録を中止することとされた。第二段階では最大 49 例の患者を追加（合計最大 62 例）し, 奏効が 14 例以上に認められた場合に有効と判定することとされた。

有効性について, 第一段階で 4 例目の奏効が認められたことから症例登録が継続され, 有効性の解析対象全例の観察期間が 6 か月以上または投与中止となった時点で解析が実施された。本試験の主要評価項目とされた RECIST ver1.1 に基づく BICR 判定による奏効率[95%CI](%)は 56.9[42.3, 70.7]であった（うち, 日本人患者 1 例（NSCLC）は PR）（表 3-4-2-1）。

表 3-4-2-1 最良総合効果, 及び奏効率(*NTRK* 融合遺伝子陽性の固形癌コホート)
(RECIST ver.1.1, 有効性の解析対象集団, BICR 判定)

最良総合効果	**例数(%)**	
	全体集団 51 例	日本人集団 1 例
CR	4(7.8)	0
PR	25(49.0)	1(100)
SD	9(17.6)	0
PD	3(5.9)	0
Non CR/PD	3(5.9)	0
Missing or unevaluable	7(13.7)	0
奏効(CR+PR)(奏効率[95%CI*](%))	29(56.9[42.3,70.7])	1(100[-,-])

-:推定不可, *: Clopper-Pearson 法

ロズリートレクカプセル審査報告書[5] 表 26 をもとに作成

審査報告書における機構の見解

機構からは真のエンドポイントであるOSと奏効率との関係は明らかではなく，STARTRK-2試験の主要評価項目とされた奏効率の結果を基に，当該患者におけるエヌトレクチニブの延命効果に関する評価を行うことは困難であること，エヌトレクチニブの有効性が検討された日本人症例数は限られていることからエヌトレクチニブの評価・結果解釈にあたり限界があるものの，癌のドライバーである *NTRK* 融合遺伝子を標的とした薬剤であるエヌトレクチニブは，癌種にかかわらず，*NTRK* 融合遺伝子陽性の進行・再発の固形癌患者に対して有効性が期待できる薬剤であると考えられることから，STARTRK-2 試験等におけるエヌトレクチニブの安全性プロファイルも考慮し，*NTRK* 融合遺伝子陽性の進行・再発の固形がん患者に対する治療選択肢として位置付けられると判断された。

3.4.3 ペムブロリズマブ (キイトルーダ®)

MSI-Highとペムブロリズマブについて[6]

マイクロサテライトはゲノム上に広く存在する反復配列であり，DNA ミスマッチ修復機構の破綻等により，DNA 複製の際に誤って複製される確率が高い配列であることが報告されている[7]。ミスマッチ修復機能が低下している固形癌はさまざまな臓器に認められ，その頻度は固形癌全体の約2~4%とされているが[8-11]，大腸がんでは6~13%[12]と癌種により異なることが知られている。

MSI-Highの表現型はミスマッチ修復関連タンパク（MLH1，MSH2，MSH6，及びPMS2）の機能欠損と密接に関係している。MSI-High を有する固形癌では，有しない固形癌と比較して体細胞変異の頻度が高く，がん抗原特異的なT細胞の標的となるネオアンチゲンが多く産生されていること，及び MSI-High を有する固形癌の腫瘍内では，活性化された細胞傷害性 T 細胞に富む微小環境が形成されていることが報告されている[13]。加えて，MSI-High を有する固形癌では，免疫を抑制性に制御するシグナル分子（PD-1 等）が高発現しており，腫瘍の排除に抵抗性の状態となっていることが報告されていること[14]を考慮すると，免疫チェックポイント阻害剤であるペムブロリズマブは，癌種を問わず，MSI-Highを有する固形癌に対して有効性が期待できると考えられた。

ペムブロリズマブの臨床試験[6]

158 試験は，大腸癌以外の症例数が少ない複数の固形癌を対象にバスケット試験として行われた国際共同第Ⅱ相試験である。有効性，及び安全性に関わる評価資料として，表 3-4-3-1 に示す国際共同第Ⅱ相試験2試験が日本における承認申請時に提出された。

表 3-4-3-1 有効性，及び安全性に関する臨床試験の一覧

資料区分	実地地域	試験名	相	対象患者	登録例数	用法・用量の概略	主な評価項目
評価	国際共同	164 試験	II	コホート A:化学療法歴のある治癒切除不能な進行・再発の dMMR または MSI-High を有する結腸・直腸癌患者	61	ペムブロリズマブ 200mg を Q3W で静脈内投与	有効性 安全性
		158 試験	II	化学療法歴のある治癒切除不能な進行・再発の MSI-High を有する固形癌患者（結腸・直腸癌を除く）	94	ペムブロリズマブ 200mg を Q3W で静脈内投与	有効性 安全性

キイトルーダ審査報告書[6] 表 30 をもとに作成

国際共同第II相試験（164 試験）は化学療法歴のある治癒切除不能な進行・再発の dMMR または MSI-High を有する結腸・直腸癌患者 61 例（日本人 7 例を含む）を対象に，ペムブロリズマブの有効性，及び安全性を検討することを目的とした単群試験のコホート A が，日本を含む 9 カ国で実施された。用法・用量は，ペムブロリズマブ 200 mg を 3 週間間隔で静脈内投与し，疾患進行または試験中止基準に該当するまで最大 35 サイクル継続することとされた。本試験の主要評価項目として，RECIST ver.1.1 に基づく中央判定による奏効率とされ，閾値奏効率は 15%と設定された。有効性について，本試験の主要評価項目とされた RECIST ver.1.1 に基づく中央判定による奏効率の結果は，表 3-4-3-2 の通りであった。なお，164 試験開示には最後に登録された患者が 6 か月観察された時点で最終解析を実施する計画であったが，なんらかの理由（審査報告書上で黒塗り）により，さらに約 6 か月追跡された時点で事後的に最終解析がされている。あらかじめ設定された最終解析時点（最後に登録された患者が 6 か月観察された時点）での奏効率の結果は，表 3-4-3-3 の通りであり，奏効率の 95%CI の下限値は，事前に設定された閾値奏効率（15.0%）を下回った。

表 3-4-3-2　164 試験における最良総合効果，及び奏効率
（RECIST ver.1.1，有効解析対象集団，中央判定）

最良総合効果	例数(%)	
	全体集団　61 例	日本人集団　7 例
CR	0	0
PR	17(27.9)	2(28.6)
SD	14(23.0)	2(28.6)
PD	28(45.9)	3(42.9)
NE	2(3.3)	0
奏効(CR+PR)(奏効率[95%CI*](%))	17(27.9[17.1,40.8])	2(28.6[3.7, 71.0])

*:正確法
キイトルーダ審査報告書[6]　表 31 をもとに作成

表 3-4-3-3　164 試験における最良総合効果，及び奏効率（最後に登録された患者が 6 か月観察された時点）（RECIST ver.1.1，有効解析対象集団，中央判定）

最良総合効果	例数(%)	
	全体集団　61 例	日本人集団　7 例
CR	0	0
PR	15(24.6)	2(28.6)
SD	16(26.2)	2(28.6)
PD	28(45.9)	3(42.9)
NE	2(3.3)	0
奏効(CR+PR)(奏効率[95%CI*](%))	15(24.6[14.5,37.3])	2(28.6[3.7,71.0])

*:正確法
キイトルーダ審査報告書[6]　表 35 をもとに作成

国際共同第II相試験（158 試験）は化学療法歴のある治癒切除不能な進行・再発の固形癌患者 94 例(日本人 7 例を含む)を対象に，ペムブロリズマブの有効性，及び安全性を検討することを目的とした単群試験が，日本を含む 15 カ国で実施された。用法・用量は，ペムブロリズマブ 200 mg を 3 週間間隔で静脈内投与し，疾患進行または試験中止基準に該当するまで最大 35 サイクル継続することとされた。また，同一の集団が安全性の解析対象とされた。MSI-High と診断された後に

本試験に登録された 83 例（グループ K）における有効性について，主要評価項目とされた RECIST ver.1.1 に基づく中央判定による奏効率の結果は，表 3-4-3-4 の通りであった。また，158 試験に組み入れられた MSI-High を有する固形癌患者の癌種別の RECIST ver1.1 に基づく中央判定による奏効率の結果は，表 3-4-3-5 の通りであった。

表 3-4-3-4 158 試験における最良総合効果，及び奏効率（グループ K）
（RECIST ver.1.1，有効解析対象集団，中央判定）

最良総合効果	例数(%)	
	全体集団 83 例	日本人集団 3 例
CR	4(4.8)	0
PR	25(30.1)	0
SD	20(24.1)	2(66.7)
PD	24(28.9)	1(33.3)
NE	10(12.0)	0
奏効(CR+PR)(奏効率[95%CI*](%))	29(34.9[24.8,46.2])	0

*:正確法

キイトルーダ審査報告書[6] 表 32 をもとに作成

表 3-4-3-5 158 試験に組み入れられた MSI-High を有する固形癌の癌腫別の奏効率（RECIST ver.1.1，有効性解析対象集団，中央判定）

癌腫	例数(%) 94 例	奏効(CR+PR) (奏効率(%))	癌腫	例数(%) 94 例	奏効(CR+PR) (奏効率(%))
子宮内膜癌	24(25.5)	13(54.2)	甲状腺癌	2(2.1)	0
胃癌	13(13.8)	6(46.2)	尿路上皮癌	2(2.1)	1(50.0)
小腸癌	13(13.8)	4(30.8)	脳腫瘍	1(1.1)	0
膵癌	10(10.6)	1(10.0)	卵巣癌	1(1.1)	0
胆道癌	9(9.6)	2(22.2)	前立腺癌	1(1.1)	0
副腎皮質癌	3(3.2)	1(33.3)	後腹膜腫瘍	1(1.1)	1(100)
中皮腫	3(3.2)	0	唾液腺癌	1(1.1)	1(100)
小細胞肺癌	3(3.2)	2(66.7)	肉腫	1(1.1)	1(100)
子宮頸癌	2(2.1)	1(50.0)	精巣腫瘍	1(1.1)	0
神経内分泌癌	2(2.1)	0	扁桃癌	1(1.1)	1(100)

キイトルーダ審査報告書[6] 表 33 をもとに作成

上記の結果を受け申請者は 164 試験において主要評価項目とされた RECIST ver.1.1 に基づく中央判定による奏効率［95%CI］(%)は 27.9［17.1, 40.8］, また, 158 試験(グループ K)において主要評価項目とされた RECIST ver.1.1 に基づく中央判定による奏効率［95%CI］(%)は 34.9［24.8, 46.2］であり, 164 試験, 及び 158 試験の対象とされた患者において臨床的に意義のある奏効率等が示された, としている。

上記を受けた, 審査報告書における機構の見解は以下であった。

治癒切除不能な進行・再発の固形癌患者における真のエンドポイントである OS と奏効との関係は明らかではなく, また, 164 試験, 及び 158 試験(グループ K)の主要評価項目とされた奏効率の結果を基に, 当該患者における本薬の延命効果に関する評価を行うことは困難であると考える。加えて, 予め設定された最終解析時点の奏効率の 95%CI の下限値は, 事前に設定された閾値奏効率(15.0%)を下回っており, 慎重に結果解釈を行う必要があると考える。しかしながら, 化学療法歴のある治癒切除不能な進行・再発の固形癌患者において, 奏効が得られることにより, 疾患進行に伴う臨床症状の改善が期待できることが報告されており[12], 当該患者において奏効が得られることは臨床的意義があると考えることから両試験の主要評価項目として奏効率を設定したという本剤の有効性に関する申請者の説明は理解可能であり, 164 試験において予め設定された最終解析時点の奏効率の結果に加えて本薬単独投与時における本薬の PK に明確な国内外差は認められていないこと, 治癒切除不能な進行・再発の結腸・直腸癌の診断, 及び治療体系に明確な国内外差は認められていないことなどの点も考慮すると, 日本人患者を含め, 化学療法歴があり, かつ他に標準的な治療のない進行・再発の MSI-High を有する固形癌患者における本薬の有効性は期待できると機構からは判断された。

ただし, 検証的な試験の成績は得られていないこと等を考慮すると, 効能・効果において本薬の投与対象が標準的な治療の適応とならない患者である旨を明記するとともに, 一次治療における有効性, 及び安全性は確立していない旨, 及び二次治療において標準的な治療が可能な場合はこれらの治療を優先する必要がある旨を注意喚起することが適切であり, 主に奏効率の結果を基に本薬の有効性の評価が行われ, 延命効果に関する情報が得られておらず, 本薬以外の治療法の実施についても慎重に検討する必要があることから, 効能・効果に関連する使用上の注意の項において本薬以外の治療法の実施を十分に考慮した上で, 本薬の投与の可否を慎重に判断する旨を注意喚起することが適切であると判断された。

3.4.4 Solanezumab, 及び gantenerumab

常染色体優性遺伝アルツハイマー病(autosomal dominant Alzeihimer's disease (ADAD))について[15, 16]

アルツハイマー病(Alzheimer's disease(AD))は, 潜行性に発症する慢性神経変性疾患であり, 認知機能や行動機能などが徐々に障害される。孤発性 AD(sporadic AD (SAD))が AD の 95%以上を占めるといわれており, 全世界で 5,000 万人以上の SAD 患者がいると推定される。発症患者は今後も増加すると予想されており, 疾患の重症度からもよりよい治療法や予防法の開発が求

められている。一方で，代替評価項目の欠如，認知機能面や臨床的な減退が非常に緩やかであること，多様な臨床表現型，認知機能などの評価の際のばらつきなどが，高い有用性を持つAD治療薬の開発の妨げとなっていると言われている。AD 予防薬については，まだ症状が現れていない状況では妥当な診断基準がないこともあり，開発の難易度は更に高くなる。

ADAD は，AD 患者のうち約 1～3%を占める，極めて稀な認知症の一つといえる。ADAD は，アミロイド前駆体タンパク質かプレセニリン（PSEN1，PSEN2）の遺伝子変異に起因する。子供への変異型を 50%の確率で伝えるとされており，遺伝子検査を通じて特定が可能である。ADAD は SAD と同様の神経病理学やバイオマーカーに関する特徴を示すが，SAD の発症は多くは 65 歳以降であるのに対して，ADAD は比較的若い年齢でも発症するとされている。

優性遺伝アルツハイマー・ネットワーク試験ユニット（DIAN-TU）について

優性遺伝性アルツハイマー・ネットワーク（The Dominantly Inherited Alzheimer Network（DIAN），https://dian.wustl.edu/）のホームページによると，DIAN は，ADAD としても知られる優性遺伝アルツハイマー病（Dominantly Inherited Alzheimer’s disease（DIAD））についての治療または予防について解決策を見つけることを目的とする国際的な研究活動団体である。Washington University School of Medicine in St. Louis によって主導されており，長期的な観察研究，基礎科学研究，及び臨床試験（治験）の計画や実施を行っている。DIAN-TU は，DIAN の臨床研究部門であり，DIAD を発症している，または危険因子のある方を対象とした介入治験のデザイン構築と管理を行っている。前述の AD 予防試験デザインにおける課題点を解決するため，企業（DIAN-TU Pharma Consortium, https://dian.wustl.edu/our-research/the-pharma-consortium/），ADAD 家族会，AD 協会，US National Institute on Aging (NIA)，規制当局（FDA，EMA），DIAN 観察研究等の研究者らのパートナーシップとして，2010 年に設立された [16]。

DIAN-TU が実施するプラットフォーム試験（NCT01760005）について [16, 17]

本試験は ADAD における予防や治療薬の特定を行うためのプラットフォーム試験としてデザインされている。2012 年，抗アミロイドベータ（Aβ）抗体である solanezumab，及び gantenerumab を評価するために，第 II/III 相，二重盲検，ランダム化，併合プラセボ対照，バイオマーカーを用いたサブ試験が最初に開始された。2014 年には，中枢神経系（Central Nervous System，CNS）の標的関与バイオマーカー（biomarker target engagement）として確立されている抗 Aβ抗体について，認知機能が正常である被験者における認知機能低下の予防を評価するための，4 年間の第 III 相アダプティブ予防試験として変更されている。本節で紹介する solanezumab，及び gantenerumab を対象としたサブ試験（NCT04623242）のほかにも，現時点では lecanemab と E2814 を評価するサブ試験が実施中である（NCT05269394）。

Solanezumab, 及び gantenerumab を評価したサブ試験について [16, 18, 19, 20]

試験デザイン

上述の通り, 本試験は, solanezumab, 及び gantenerumab について, 4 年間の認知機能に関する評価した多施設共同, 二重盲検, ランダム化, 併合プラセボ対照試験である。ADAD のリスクまたは軽度の症候性 ADAD を持つ変異保持者, すなわち, 臨床的認知症尺度（clinical dementia rating, CDR）が 0～1 でかつ推定発症年齢からの年数（estimated years from symptom onset, EYO)が-15 年から+10 年までの ADAD 変異保持者を事前に計画された評価対象としている[16]。推定発症年齢からの年数（estimated years from symptom onset, EYO)とは, 各被験者が試験の評価を受ける時点での年齢と同被験者の ADAD 発症の推定年齢の差として定義している。例えば, ある被験者がある評価を受けた時点の年齢が 45.3 歳, その症例で推定発症年齢が 50.2 歳とすると, 評価時点での EYO は-4.9 年となる。使用する推定発症年齢には, システマティック・レビューとメタ・アナリシスに基づきマッチングされる集団の発症年齢の平均値が用いられる[19, 20]。

変異がない被験者にはプラセボが割り当てられ, 事前に計画された解析には含まれない。変異のある被験者には, 実薬かプラセボが 3:1 にランダム化され, 割り付け結果については被験者, 実施者, スポンサーのいずれも盲検化される。変異のある症例数としては, 計画時には 160 例（solanezumab vs プラセボでそれぞれ 60 例, 20 例。Gantenerumab vs プラセボでそれぞれ 60 例, 20 例。）を予定していた。Solanezumab とのランダム化でプラセボに割り当てられた被験者と, gantenerumab とのランダム化でプラセボに割り当てられた被験者は併合プラセボ群として, 解析の際の比較対照に用いられた[18]。

Solanezumab, gantenerumab の投与量についてはそれぞれの Phase I/II 試験におけるバイオマーカー等の結果から試験開始時に設定するものの, 非盲検チームによる安全性やバイオマーカーの中間解析結果によって, 用量の変更も可能とする計画であった[16]。

主要評価項目は DIAN 多変量認知評価項目（DIAN Multivariate Cognitive End Point, DIAN-MCE)であり, その他, 複数の認知及び臨床評価, 画像バイオマーカーなどを含む副次評価項目が設定されている。DIAN-MCE には, Wechsler Memory Scale-Revised Logical Memory, Wechsler Adult Intelligence Scale Digit Symbol Substitution Test (Digit Symbol), International Shopping List Test (ISLT) Delayed Recall score, 及び Mini-Mental State Examination (MMSE)が含まれ, それぞれの z-スコアの平均値を複合評価項目として取り扱う[16, 18]。

解析にはベイズ流多変量認知機能増悪モデル （Bayesian multivariate cognitive disease progression model （DPM））を用いた。DPM の説明のため, 初めに治療介入前における認知機能低下のモデル化を簡単に説明する。今, $Y_{ij}, i = 1, \cdots, k, j = 1, \cdots, n_i$を被験者$i$における$j$番目の測定値（複合評価項目）とし, 対応する EYO（年）をEYO_{ij}と表記する。この時, DPM として以下の混合効果モデルを仮定した。

$$Y_{ij} = \gamma_i + f\left(EYO_{ij} + \delta_i \middle| \boldsymbol{\alpha}\right) + \varepsilon_{ij}, \boldsymbol{\alpha} = (\alpha_{-15}, \alpha_{-14}, \cdots, \alpha_{15})^T$$

γ_iは個々の被験者における比較的健康な状態（EYO が－15 以下）における複合評価項目の値を示すランダム効果であり, δ_iは発症年齢についての不確定さを考慮したランダム効果である。$\boldsymbol{\alpha}$は

関数fのパラメータであり，関数fには以下の単調減少を前提とするスプライン関数を用いた。

$$f(x)=\begin{cases}0, & (x\le -15)\\ (1+\lfloor x\rfloor -x)\alpha_{\lfloor x\rfloor}+(x-\lfloor x\rfloor)\alpha_{\lfloor x\rfloor+1}, & (-15<x\le 15)\\ \alpha_{15}, & (x>15)\end{cases}$$

ただし，$\lfloor x\rfloor$は床関数（x未満の最大の整数を戻り値とする）であり，このモデルの下では関数fはEYOのカテゴリごとで線形に低下する区分線形関数として表される。ただし，治療開始以降は治療による影響を受けるため，上記の定式化については拡張が必要となる。今，T_i を被験者i における治療開始の時期（EYO と同じく推定発症年齢との差として算出）とし，$Y_{ij}, i=1,\cdots,k, j=1,\cdots,n_i$についてのモデルを以下のように拡張する。

$$Y_{ij}=\begin{cases}\gamma_i+f\left(EYO_{ij}+\delta_i\middle|\boldsymbol{\alpha}\right)+\varepsilon_{ij}, & (EYO_{ij}<T_i)\\ \gamma_i+f(T_i+\delta_i|\boldsymbol{\alpha})+e^{\theta}\left[f\left(EYO_{ij}+\delta_i\middle|\boldsymbol{\alpha}\right)-f(T_i+\delta_i|\boldsymbol{\alpha})\right]+\varepsilon_{ij}, & (EYO_{ij}\ge T_i)\end{cases}$$

この時，θは治療効果に関するパラメータであり，e^{θ}は新規治療のプラセボに対する相対的な効果の大きさを表す cognitive progression rate（CPR）と考えられる。例えば，CPR が1.00であれば認知機能低下の速度はプラセボ群と同じ，CPR が0.70であれば，治療によってプラセボ群よりも30%認知機能低下の進展が遅らせられると解釈できる。本試験では，CPR が 1 未満となる事後確率が0.981を超えた場合に，効果があると判定することとした[18, 19]。

試験結果 [18]

本サブ試験では，ランダム化され，かつ二重盲検期において1回以上の試験治療を受けた変異のある被験者は 144 例であった。このうち，gantenerumab，及び gantenerumab プラセボが投与されたのはそれぞれ 52 例，及び 21 例であり，solanezumab 及び solanezumab プラセボが投与されたのはそれぞれ 52 例，及び 19 例であった。なお，各プラセボ群については，解析上は併合して solanezumab 群，及び gantenerumab 群との比較に用いた。背景因子について，gantenerumab 群，solanezumab 群，及び併合プラセボ群間での特筆すべき不均衡は認められなかった。

有効性の結果については以下の通りである。併合プラセボ群に対する gantenerumab 群の CPR について，事後分布の平均値は 1.063 であり，CPR が 1 未満となる確率は 0.144 であった。また，併合プラセボ群に対する solanezumab 群の CPR について，事後分布の平均値は 1.255 であり，CPR が 1 未満となる確率は 0.0001 であった。事前に規定した，CPR が 1 未満となる事後確率が 0.981 超という基準を満たさなかった。

なお，参照した文献[18]では，解析に使用したモデルの仮定が実際に試験で得られた結果では満たされていなかった点について考察している。解析モデルでは，EYO の値に応じ，症候性・無症候性の被験者における単調な認知機能の減弱や同じ分散を仮定しているが，実際の結果では，無症候性（CDR=0）の被験者の集団において認知機能の減弱が認められず，むしろ同集団ではコンポーネント 1 つである Logical Memory Delayed Recall Test で学習効果の存在が示唆され，その他の評価項目でも経時的な改善が認められた。しかしながら，主要評価項目や CDR-Sum of Boxes（CDR-SB），Functional Assessment Scale（FAS）といった副次評価項目についての反復測

定混合モデルによる探索的な感度解析でも，上述の結論は変わらなかったことが報告されている。いくつかのバイオマーカーについては，gantenerumab 群では併合プラセボ群と比較して改善する傾向が示唆され，特に疾患が早期の患者層においてアミロイド低下による生物学的な進行の予防や鈍化が起こる可能性も考えられたが，いずれの薬剤も認知機能や臨床的なベネフィットについては示唆されなかったとまとめられている[18]。

まとめ

3.4 節では，マスタープロトコルを用いた試験デザインの概要を示し，またいくつかの試験デザインや用いられている解析手法について紹介した。事例としては，3.4.2，3.4.3 節では日本で承認申請が行われたエヌトレクチニブとペムブロリズマブを挙げた。バイオマーカーを持った患者集団を対象とし，がん腫横断的な治療効果を検討するバスケット試験では，これらの事例のようにがん種を併合した主要解析が実施されるケースが考えられる。その場合はがん種横断的な併合を行う妥当性を非臨床・臨床的な観点から説明する必要があるであろう。3.4.4 節では，承認申請が行われた事例ではないものの，参考として Washington University School of Medicine が実施する ADAD を対象としたプラットフォーム試験において公表されているサブ試験を紹介した。当該サブ試験では，患者層を考慮した新たな評価項目の設定，推定発症年齢を考慮した解析モデルの適応，プラセボ比較の際の被験者共有（併合プラセボ群との比較解析）など，より効率的なエビデンス構築に向けた取り組みが多数組み込まれていた。更に同じ組織（DIAN-TU）の下で試験を実施できることで，被験者組み入れや評価など試験運営が治療間で類似すること，実施における効率性，さらに今後も併用治療も含む新しいサブ試験の追加も可能であるなど，プラットフォーム試験を用いることの長所がいくつも考えられた。また，DIAN では，DIAN 観察研究や DIAN 拡大レジストリなどを通した ADAD 患者とのエンゲージメントがあり，医師や施設から紹介される潜在患者数は 3500 人超といわれている[2]。ADAD のような希少疾病では，患者特定やヒストリカルコントロールとしての活用の可能性など，こういったネットワークを活用することの大きなメリットとも考えられ，上述のプラットフォーム試験の実施も含め，DIAN-TU の取り組みは，こういった希少疾病医薬品開発におけるチャレンジを克服するための 1 つのモデルとも考えられた。

希少疾病用医薬品の開発では，患者数が少ないことや倫理的な面からプラセボ同時対照が設定できないなどの理由から，十分な検出力の下で治療効果を精度高く推測することが困難な場合がある。マスタープロトコルを用いた臨床試験では，その目的やデザインによって，患者さんが治験にアクセスするハードルを下げる（例えば，バイオマーカーの活用によって被験者がより効果の高い治療に割り当てられる可能性が高まる，または対照群への割り付け比率を低く設定する等），本節で述べたインフラストラクチャの面での便益をもたらす，革新的なデザインや解析手法による効率的なエビデンス構築やデータに基づく意思決定が行われる，といった点が期待され，国内外での更なる活発な議論が今後も望まれる。

参考文献

[1] A Drilon, et al. ROS1 dependent cancers-biology, diagnositcs and therapeutics. Nat Rev Clin Oncol 18.1(2020): 35-55.

[2] M Soda, et al. "Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer." Nature 448.7153 (2007): 561-566.

[3] E Cocco, M Scaltriti, A Drilon. "NTRK fusion-positive cancers and TRK inhibitor therapy." Nature reviews Clinical oncology 15.12 (2018): 731-747.

[4] AM Lange, HW Lo. "Inhibiting TRK proteins in clinical cancer therapy." Cancers 10.4 (2018): 105.

[5] 医薬品医療機器総合機構. "ロズリートレクカプセル，審査報告書."（2019）. https://www.pmda.go.jp/drugs/2019/P20190716001/450045000_30100AMX00015_A100_2.pdf

[6] 医薬品医療機器総合機構. "キイトルーダ，審査報告書." （2018）. https://www.pmda.go.jp/drugs/2019/P20190115001/170050000_22800AMX00696000_A100_1.pdf

[7] GM Li. "Mechanisms and functions of DNA mismatch repair." Cell research 18.1 (2008): 85-98.

[8] A Latham, et al. "Microsatellite instability is associated with the presence of Lynch syndrome pan-cancer." Journal of clinical oncology 37.4 (2019): 286-295.

[9] Cortes-Ciriano, Isidro, et al. "A molecular portrait of microsatellite instability across multiple cancers." Nature communications 8.1 (2017): 1-12.

[10] Bonneville, Russell, et al. "Landscape of microsatellite instability across 39 cancer types." JCO precision oncology 1 (2017): 1-15.

[11] Steiniche T, et al. "5252 - Analytic Validation of Tumor Mutational Burden as a Companion Diagnostic for Combination Immunotherapy in Non-Small Cell Lung Cancer." Annals of Oncology (2018) 29 (suppl_8): viii14-viii57. 10.1093/annonc/mdy269

[12] Le, Dung T., et al. "PD-1 blockade in tumors with mismatch-repair deficiency." New England Journal of Medicine 372.26 (2015): 2509-2520.

[13] Giannakis, Marios, et al. "Genomic correlates of immune-cell infiltrates in colorectal carcinoma." Cell reports 15.4 (2016): 857-865.

[14] Llosa, Nicolas J., et al. "The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints." Cancer discovery 5.1 (2015): 43-51.

[15] Imbimbo, Bruno P., et al. "Accelerating Alzheimer's disease drug discovery and development: what's the way forward?." Expert Opinion on Drug Discovery 16.7 (2021): 727-735.

[16] Bateman, Randall J., et al. "The DIAN-TU Next Generation Alzheimer's prevention trial: adaptive design and disease progression model." Alzheimer's & Dementia 13.1 (2017): 8-19.

[17] Kidwell, Kelley M., et al. "Application of Bayesian methods to accelerate rare disease drug development: scopes and hurdles." Orphanet journal of rare diseases 17.1 (2022): 1-15.

[18] Salloway, Stephen, et al. "A trial of gantenerumab or solanezumab in dominantly inherited Alzheimer's disease." Nature medicine 27.7 (2021): 1187-1196.

[19] Wang, Guoqiao, et al. "A novel cognitive disease progression model for clinical trials in autosomal‐dominant Alzheimer's disease." Statistics in medicine 37.21 (2018): 3047-3055.

[20] Ryman, Davis C., et al. "Symptom onset in autosomal dominant Alzheimer disease: a systematic review and meta-analysis." Neurology 83.3 (2014): 253-260.

3.5 小児での有効性に関する成人データの借用

3.5.1 概要

Therapeutic orphan（治療上の見捨てられた孤児）という言葉があるように，小児等において，医薬品の十分な評価が行われておらず，適応外使用がされている状況は，以前から問題点として指摘をされてきた[1]。FDA の Rare Diseases: Common Issues in Drug Development Guidance for Industry [2]では希少疾病患者の半数は小児であるとされており，小児における有効性，及び安全性を評価することは多くの希少疾病において重要であることが述べられている。

小児を対象とした医薬品開発を行う上での留意点として，まず小児では成人と比較すると対象症例数が少ない場合が多い[3]。ランダム化比較試験を行う際には，自分の子供がランダムに治療群へ割付されることへ懸念を示す保護者も多いことが知られており，症例登録が難しい場合がある。小児は成人よりも脆弱な集団であるので臨床試験を行う場合には最低限の症例数で試験を実施するべきである。しかし，通常の試験デザインや解析手法を用いるとこれらの臨床試験に必要な症例数は，成人のそれと同じになってしまう。そこで症例数の制約や，有効性に対する規制要件を満たすためのアプローチとして成人で得られたエビデンスの小児への外挿が知られている[4]。

ICH E11 ガイドライン[5]では成人で試験され承認されたものと同じ適応症を対象とした医薬品が小児に使用される場合，疾患経過が成人と小児で類似しており，治療結果の比較が可能であると推定できるのであれば，成人における有効性データを外挿することが可能であると記載がされている。また ICH E11 R1 ガイドライン[6]では小児用医薬品開発における外挿を，「疾患経過，及び期待される医薬品への反応が，小児，及び参照集団（成人または他の小児集団）の間で十分に類似していると推定できる場合に，小児集団における医薬品の有効かつ安全な使用を支持するエビデンスを提供する手段」としてより明確に定義し，そのプロセスについて述べている。また小児用医薬品開発における外挿に関する包括的な枠組みを提供する ICH E11A が現在作成されている。

一方，小児と参照集団の類似性に不確実性がある場合には，有効性に関する試験を小児でも実施し，成人での有効性のデータを小児で借用するアプローチも考えられる[5]。限りのある小児のデータに，成人のデータを加えることで統計的に精度の高い推定が可能となる。しかし，小児と成人でプロファイルが異なる場合に，多くの情報を借用すると誤った結論を得る可能性が高くなることに留意する必要がある。EMA の小児での外挿に関する Reflection paper [7]では，成人から小児への外挿が妥当と判断された場合に考えられる方法として，5%より大きい有意水準，大きな非劣性マージンの使用とともにベイズ流の方法を紹介している。外部データ（成人データ）を用いることに対する不確実性の程度は，どの程度小児試験で症例数を減らすかに反映されるべきであり，ベイズ流の解析で事前分布として含められる情報量は，それまでに得られているエビデンスの頑健性によってケースバイケースで決まることが記載されている。情報借用を行う場合は臨床試験で得られるデータと比べて，どの程度の事前情報が用いられるかを定量化することが重要であること，ベイズ流の解析を実施した場合に頻度流でいう第一種の過誤確率に関する性質も調べておく必要があると述べられている。

ここでは成人から小児への有効性データの外挿された結果に基づき，FDA と議論がされた事例

としてベリムマブの全身性エリテマトーデスでの議論を紹介する [8, 9]。

3.5.2 ベリムマブ（ベンリスタ®）

小児における全身性エリテマトーデス（SLE）について

全身性エリテマトーデス（SLE）は自己免疫疾患であり，発熱や全身倦怠感のほかに皮膚症状，関節炎などの様々な症状が全身に現れる。有病率は 20~70 人/10 万人であり，女性が多いことが知られている。15~40 歳での初発が多いが，10~20%は 20 歳までに発症し，その有病率は世界で 4.3~9.73 人/10 万人，特に 9 歳以下での症例は非常に少ないことが知られている。小児発症と成人発症ではおおむね症状は同様であるが，小児のほうが腎機能や精神神経に関する症状が起こることが知られている。ベリムマブ承認以前は小児発症の SLE に対して承認された薬剤はなく，成人で承認されている免疫抑制剤の組み合わせで治療されることが多かった。

ベリムマブについて

ベリムマブは完全ヒト型抗 BLyS モノクローナル抗体製剤である。B 細胞に「可溶性 B リンパ球刺激因子（BLyS）」と呼ばれる因子が結合することで，B 細胞の活性化・生存，形質細胞へ分化するが，これが SLE の患者では過剰発現していることで知られている。ベリムマブは BLyS を阻害することで，B 細胞の生存抑制や形質細胞への分化を抑制し，結果的に抗核抗体の産生を抑制すると考えられている。米国においてベリムマブは 2011 年に点滴静注製剤が活動型自己抗体陽性の SLE に対して承認された。

PLUTO 試験

小児全身性エリテマトーデス（SLE）患者にベリムマブ 10mg/kg を 48 週間静脈内投与したときの有効性，安全性，忍容性，薬物動態，バイオマーカー，及び生活の質（QOL）への影響を評価するために PLUTO 試験が実施された。本試験はパート A, B, 及び C の 3 つのパートで構成されていた。パート A 内は 52 週のプラセボ対照の二重盲検比較試験であり，各年齢層で安全性と薬物動態を確認するため，コホート 1 から 3 で構成されている。コホート 1 は 12～17 歳，コホート 2 は 5～11 歳，コホート 3 は 5～17 歳の患者が対象となる。コホート 1 は 12 例，コホート 2 は 10 例以上，コホート 3 は 48 例以上の患者で構成され，コホート 1 及び 2 はいずれもベリムマブと標準治療の併用群とプラセボと標準治療の併用群へ 5:1 の割付比，コホート 3 は 1:1 の割付比で割付がされた。コホート 1 で最低 8 週の投与終了時に安全性と薬物動態に関する解析を実施し，その結果を受けてコホート 2，及びコホート 3（ただしコホート 3 は 12～17 歳のみ）の登録を開始した。同じくコホート 2 で最低 8 週の投与終了時に安全性と薬物動態に関する解析を実施し，その結果を受けてコホート 3 の 5～11 歳患者の登録を開始するデザインであったが，コホート 3 の 5～11 歳患者の登録開始前に目標例数が達成されたため，実際には組み入れられなかった。パート B はパート A を完了した患者での長期非盲検下での安全性の追跡，パート C はパート A または B を中止した患者での長期安全性の追跡を行う。有効性の評価に関してはパート A で行われるため，以降は主にパ

ートAについて述べる。主要評価項目は52週時点のSLE Responder Index（SRI）の反応率であった。主要解析はプラセボに対するオッズ比とその95%信頼区間の算出であった。小児SLEの希少性から，症例数設定は検出力に基づいたものではなく，精度ベースの設定であった。

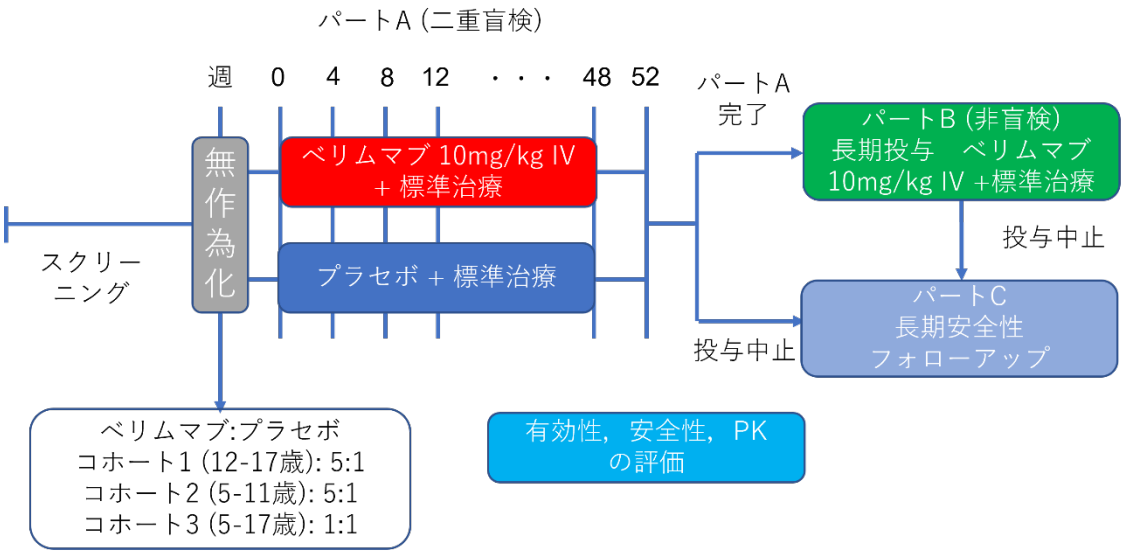


図 3-5-2-1 PLUTO 試験デザイン

Brunner et al. (2020) [10] をもとに作成

試験結果

PLUTO試験ではベリムマブ群に53例，プラセボ群に40例が主要解析に含まれ，反応率はベリムマブ群で52.8%，プラセボ群で43.6%，反応率のプラセボ群に対するオッズ比は1.5 [95%信頼区間 0.6-3.5]であり，本試験では仮説検定に基づく症例数設計はされていなかったものの，信頼区間の下限は1を下回っていた。このオッズ比の点推定値は後述する2つの成人試験でのそれと類似していたが，小児試験では例数が少なく信頼区間は広かった。また薬物動態に関しても成人と小児で類似した結果が得られていると結論づけられている。

成人データの借用

FDAのClinical Review teamは疾患及び治療の奏効は成人でも小児でも変わらないはずであると考えた。PLUTO試験は成人を対象に実施された試験と同じ主要評価項目を用いていたこと，適格基準，製剤，投与方法や投与量は2つの成人試験と類似していたことから，FDAは過去の成人試験（1056, 1057 試験）の情報を借用したベイズ流の解析の実施を照会事項対応として要求した。また具体的な手法としてStatistical review teamは以下の解析の要求をした。

表 3-5-2-1 ベリムマブ成人試験の SRI 有効率の結果

	1056 試験			1057 試験		
	プラセボ N=275	ベリムマブ 1mg/kg N=271	ベリムマブ 10mg/kg N=273	プラセボ N=287	ベリムマブ 1mg/kg N=288	ベリムマブ 10mg/kg N=290
有効例 例数 (%)	93 (34)	110 (41)	118 (43)	125 (44)	148 (51)	167 (58)
群間差		7%	9%		8%	14%
オッズ比 (95%CI)		1.3 (0.9, 1.9)	1.5 (1.1, 2.1)		1.6 (1.1, 2.2)	1.8 (1.3, 2.6)

FDA Multi-disciplinary Review をもとに作成[8]

a 成人の試験をメタ・アナリシスすることで治療効果に関する単一の推定値，及び分布を得ること

b Skeptical prior（懐疑的な事前分布）を小児で設定すること。例えば平均を 0 とし SD を過去の試験から得た正規分布から設定することが考えられる。

c 上記の二つの要素を重みづけした小児試験の事前分布を設定する。具体的には$f(D)$を skeptical prior, $g(D)$を成人の治療効果の分布，a を成人データの重みとした際に以下の小児事前分布を仮定する。

$$\text{小児での事前分布 } = (1-a) \times f(D) + a \times g(D)$$

d 事後分布で治療効果が 0 を超える確率として，有効性の事後確率を算出する。

e 上記の重みa を 0 から1まで 0.05 ずつ変えて実施する。

f 有効性の事後確率がある高い閾値（95%, 97.5%, 99%）を超えるa を決定する。

g 各重みでの有効性の事後分布を図示する。

この要求に対してスポンサーは以下の流れでベイズ流の解析を実施した。

まずy_p，及びy_Aを観測された小児（成人）での対数オッズ比，s_p^2 ，及びs_A^2を小児（成人）での対数オッズ比の分散の推定値，δ_pを小児での治療効果のパラメータ，m を小児での各群 1 例に対応する有効サンプルサイズ [4]，aは小児での事前分布を作成するうえで成人データの重みである。なお，成人のデータは 2 試験の結果を併合して利用している。この時，小児での治療効果に関する事前分布は

$$\delta_p \sim (1-a) \times N\left(0, m * s_p^2\right) + a \times N(y_A, s_A^2)$$

と記載できる。事前分布の第一項が skeptical prior に対応しており，第二項の成人データから導出

4 有効サンプルサイズとはベイズ流の解析をする上で，事前分布が持っている情報量である。例えば，n個の標本で二項分布に従う変数Yに対して，ベータ分布に従う事前分布 Beta(a, b)を用いた場合，その事後分布は Beta(a+Y, b + n - Y)となる。つまりn個の標本から得られたデータを加えることによって，事前の合計$a+b$が事後の合計$a+b+n$になったと考えることができるので，事前分布の情報量，すなわち有効サンプルサイズは$a+b$と考えることができる。[9]

される事前分布と重みa で併合されているのがわかる。小児試験での標準偏差を SD，N_pを小児試験の総症例数(= 92)とした際に，$s_p^2 = SD^2/N_p$を仮定する。また，m を小児での各群 1 例（計 2 例）での有効サンプルサイズとするので，$m * s_p^2 = SD^2/2$ となるように m を設定する。この場合，

$$m * s_p^2 = SD^2/2 \rightarrow m\frac{SD^2}{N_p} = \frac{SD^2}{2} \rightarrow m = \frac{N_p}{2}$$

からmが 46 と決まる。成人試験及び小児試験（PLUTO 試験）のデータより，$s_A = 0.121$及び$s_p = 0.424$が得られたため，最終的に小児での事前分布の式は以下のように求められる。

$$\delta_p \sim (1-a) \times N(0,\ 8.27) + a \times N(0.48, 0.015)$$

事前分布の形状を成人データの重みa を変えた際にどうなるかを以下の図 3-5-2-2 に示す。成人の情報を借用するほど，事前分布の形状がより鋭くなることがわかる。

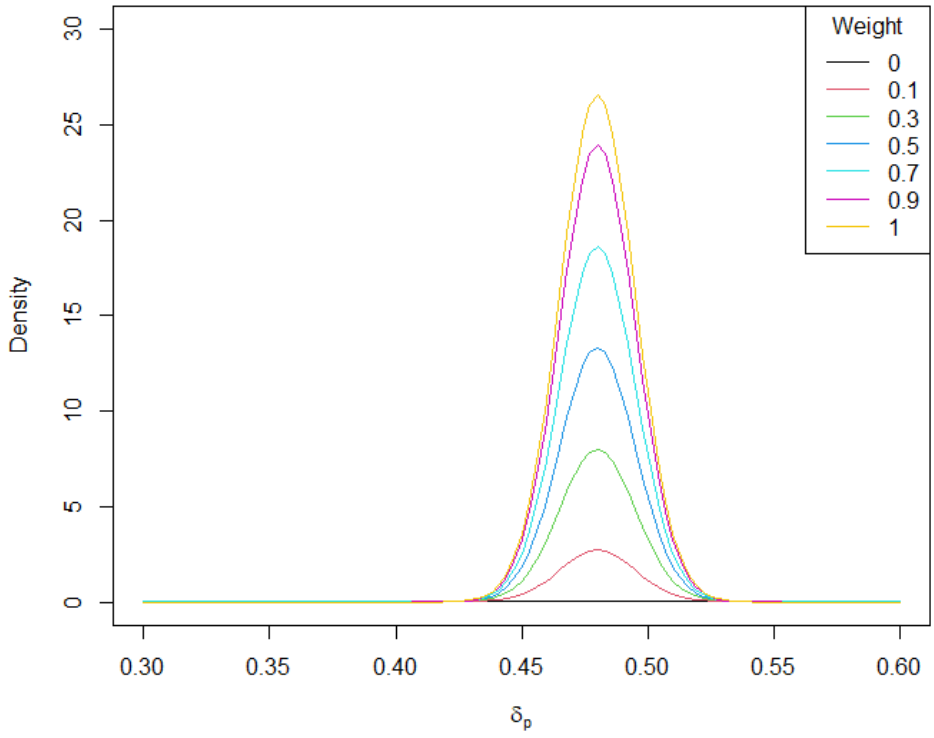


図 3-5-2-2 異なる重みを使った場合の小児試験における事前分布
FDA Multi-disciplinary Review をもとに作成[8]

この事前分布に基づき，小児試験で得られた結果も加味した事後分布，及び 95%信用区間が算出された。事後分布は先述した事前分布における成人データの重みごとに計算された。その結果を図 3-5-2-3 に示す。横軸は成人データの重みa を示し，縦軸はa を変えた場合のオッズ比と

その 95%信用区間である。成人データの重みが増大するに従い，事前分布，及び事後分布で用いる成人データの情報量が多くなり，信用区間が狭くなっているのがわかる。95%信用区間は重みが 0 の時には 1 を含んでいるが，重みが 0.55 を超えたところから 1 を含まなくなっている。これは有効性がある事後確率が 97.5%以上であることに対応している。

FDA はこの結果から以下のように考察している。まず小児試験の有効性の結果は，症例数が限られていることから信用区間が広くなっているが，過去に実施された成人試験の結果と類似しており，情報借用が合理的と考えられることが述べられている。またおおよその背景情報も成人と小児で類似していることも述べられている。さらに臨床チームとの議論の結果，本疾患で成人と小児での 55%の情報の関連度とすることは合理的と考え，ベリムマブは小児集団でも最低 97.5%の事後確率で有効性があると考えると結論付けている。

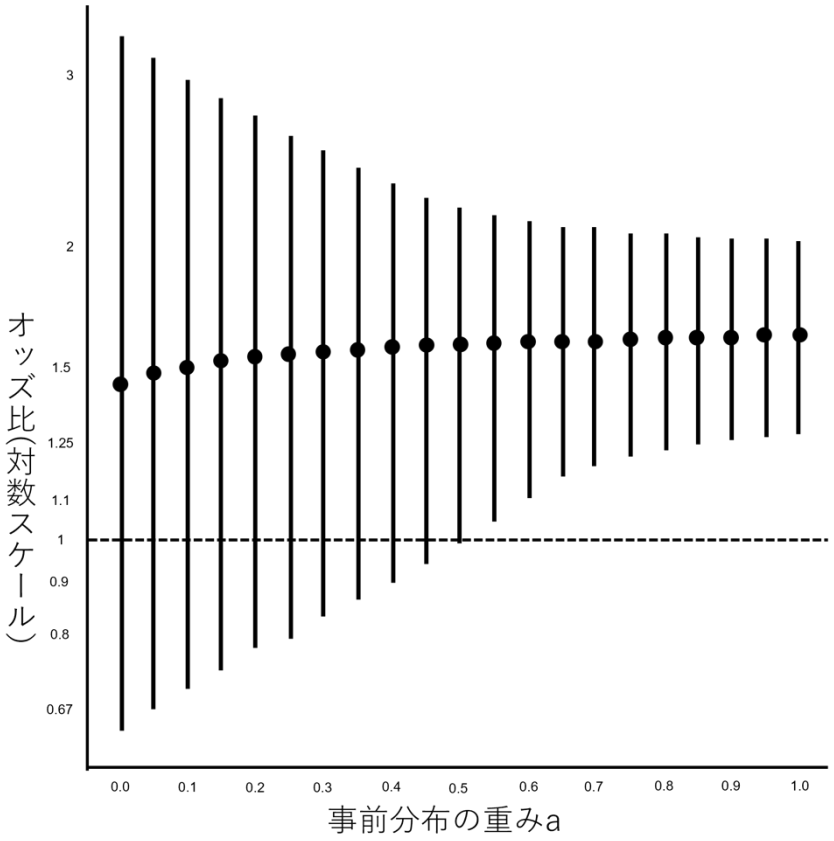


図 3-5-2-3 SRI レスポンスに関するオッズ比の事後分布の平均と 95%信用区間
FDA Multi-disciplinary Review をもとに作成[8]

今回の結果では，臨床的な観点から 55%の情報借用が合理的と考えられているが，どの程度までの情報借用が受け入れ可能であるかは FDA Multi-disciplinary Review では記載されていない[8]。例えば 80%情報借用しないと信用区間の下限が 1 を超えない場合はどうなるのかは不明である。またこの借用の受け入れ可能性は疾患，薬剤，試験デザインなどによって変化すると考えられる。

なお FDA の Statistical reviewer は追加で事後分布の算出にベースラインや年齢グループを加味した事後分布の算出を実施している。また本試験に関する原著論文[10]の Supplementary material 2 には副次評価項目である SELENA SLEDAI に関するベイズ流の解析結果も記載されている。

なお，この議論は照会事項対応での FDA とのやり取りである。日本の審査報告書では成人データからの情報借用については触れておらず，小児 SLE 患者が少ないことから検証的試験が困難なこと，成人試験と同様のデザインとした小児試験の結果から有効性・安全性の評価は可能とし，承認された[11]。

まとめ

本節では小児を対象とした開発で，対象集団が小さく，統計的に頑健な結果を得ることが困難な場合に，成人データの借用を行った事例を紹介した。今回の事例は照会事項の中での適用事例であったが，臨床試験の計画時に成人データの借用を計画することで，小児での症例数を抑えることが可能な場合もあると考える。しかしどの程度の情報借用が可能かは，疾患や薬剤の特性に応じてケースバイケースと考えられる。今回紹介したように事後的に情報借用をした事例はあるものの，事前に規制当局と合意できることが望ましいであろう。

参考文献

[1] Shirkey, Harry C. "Therapeutic Orphans—Everybody's Business." Drug Intelligence 2.12 (1968): 323-323.

[2] Food and Drug Administration. "Rare Diseases: Natural History Studies for Drug Development - Guidance for Industry (Draft)." (2019).

[3] 坂口宏志, 石川資子, and 崎山美知代. "小児製剤に関する欧米の取り組みと日本における今後の課題." 薬剤学 75.1 (2015): 5-8.

[4] Gamalo-Siebers, Margaret, et al. "Statistical modeling for Bayesian extrapolation of adult clinical trial information in pediatric drug evaluation." Pharmaceutical Statistics 16.4 (2017): 232-249.

[5] 厚生労働省. "小児集団における医薬品の臨床試験に関するガイダンス." (2000).

[6] 厚生労働省. "「小児集団における医薬品開発の臨床試験に関するガイダンス」の補遺." (2017).

[7] European Medicines Agency. "Reflection paper on the use of extrapolation in the development of medicines for paediatrics." (2018).

[8] Food and Drug Administration. "Multi-disciplinary Review and Evaluation Benlysta® (belimumab) for Intravenous Infusion in Children 5 to 17 Years of Age with SLE." (2019) https://www.fda.gov/media/127912/download

[9] Morita, Satoshi, Peter F. Thall, and Peter Müller. "Determining the effective sample size of a parametric prior." Biometrics 64.2 (2008): 595-602.

[10] Brunner, Hermine I., et al. "Safety and efficacy of intravenous belimumab in children with systemic lupus erythematosus: results from a randomised, placebo-controlled trial." Annals of the rheumatic diseases 79.10 (2020): 1340-1348.

[11] 医薬品医療機器総合機構. "ベンリスタ点滴静注用, 審査報告書." （2019）. https://www.pmda.go.jp/drugs/2019/P20190917002/340278000_22900AMX00985_A100_1.pdf

3.6 試験デザインと解析手法についてのその他の議論

3.6.1 CID パイロットプログラムの中で実施された試験

3.6.1.1 概要

Complex Innovative trial design（CID）のパイロットプログラムは FDA が 2018 年 8 月に開始したプログラムである[1]。ここでは，FDA ガイダンス Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products [2]に基づき，CID について簡単に紹介をする。

CID は，複雑なアダプティブ，ベイズ流，及びその他の新規の臨床試験デザインを指すと考えられているが，革新的または新規と考えられるものは時間とともに変化する可能性があるため，CID の固定された定義はない。CID には新薬承認申請または生物学的製剤承認申請における有効性の証拠を提供する目的で，これまでほとんどまたは一度も用いられたことのない試験デザインが含まれ，試験の動作特性を推定するためにシミュレーションが必要なことが特徴である。そのため目的，実施，動作特性（誤った結論を導く可能性など）を含むデザインの側面，及び試験データをどのように解析し提示するかについて，治験依頼者と FDA との間で明確なコミュニケーションが求められる。CID を提案する場合の例として，以下に該当する臨床試験ではその内容を提案に含める必要がある。

- 試験デザインの選択，及び医薬品開発計画全体にどのように適合するかについての考察
- 試験デザインの重要な側面（アダプテーションの計画，中間解析の詳細，決定基準を含む）に関する詳細な説明
- CID に関連した統計的な考察
- 事前情報を用いる場合は，事前情報の情報源及び選択方法の詳細，妥当性及び関連がある全ての事前情報が考慮されていることを確認するための手順に関する説明
- デザインの動作特性の詳細な評価（誤った結論を導く可能性，及び治療効果の推定の信頼性を含む。ベイズ流の推測を用いる場合は，事前分布の選択の影響）
- シミュレーションが実施された場合は，シミュレーションに関する報告書
- データモニタリング委員会，または CID の重要な側面を担う組織についての役割や指示書
- データへのアクセスを適切に制限し，試験の完全性を維持するための包括的な計画
- 患者の意見を試験デザインや解析に取り入れる場合には，患者の意見を収集するための試験計画書

さらにベイズ流解析手法を用いた CID を提案する場合には以下の二つの要素について説明することが推奨されている。

<u>事前分布</u>

一般に，ベイズ流の CID の提案には事前分布に関する詳細な考察を含める。事前分布の作成に用いたデータまたはその他の外部情報を詳細に示し，FDA が外部情報の情報源，及び完全性，その妥当性，並びにデータの質，及び信頼性を理解できるようにする。外部データの妥当性は，交

換可能性の問題に関連しており，これはベイズ流の CID の提案において言及されるべきである。

決定基準

治験依頼者は，添付文書に含めることを意図した主要評価項目，及び副次評価項目について，治験実施計画書において，決定基準（例：$\Pr(\pi_A > \pi_B) > 0.99$）を提案する必要がある。また，決定基準の設定根拠を含める必要がある。

こうした CID は，例えば集団の規模が小さい疾患領域またはアンメットメディカルニーズがある領域で，従来の医薬品開発が困難または最適でない場合に，特に有望となり得る[3]。特に，母集団が小さい場合には，症例数を低減できる革新的な試験デザインは，開発を加速するだけでなく，実施不可能な開発プログラムを実施可能にすると考えられており，希少疾病ではこういったデザインを検討することが有用な場合もあると考える。

Price and Scott (2021) [4]では，執筆された時点で CID プログラムに採択された 5 つの試験が，簡潔に紹介されている。選択された試験の疾患領域は，デュシェンヌ型筋ジストロフィー，小児多発性硬化症，慢性疼痛，全身性エリテマトーデス，アンメットメディカルニーズのある腫瘍領域（詳細は記載されていない）であり，デュシェンヌ型筋ジストロフィー，小児多発性硬化症は希少疾病である。CID を行うモチベーションは多岐にわたるが，これらの 5 つの試験ではいずれもベイズ流の解析手法が使用されているのが特徴である。一方，CID パイロットプログラムに採択されなかった事例も存在するが，その理由は疾患領域における適切な主要評価項目が明確になっていないこと，進行中の試験において革新的な試験デザインは含まれておらずむしろ主要評価項目の改訂が行われていること，関連する審査部門から既に様々な助言を受けていることが判明したことであった。これらに関しては既存のルートを通じて規制当局のフィードバックを求めることが推奨された。

CID プログラム自体が比較的最近始まったこともあり，本報告書の執筆時点で，このプログラムを通じて承認されている薬剤はなく，試験デザインに関する報告は限られているが，デュシェンヌ型筋ジストロフィーを対象にした suvodirsen の試験デザインが報告されているため，以下に紹介する[5]。

3.6.1.2 suvodirsen

デュシェンヌ型筋ジストロフィーについて

デュシェンヌ型筋ジストロフィー（DMD）に関する説明は 3.1.5.3 節を参照していただきたい。後述する DYSTANCE 51 試験計画時の DMD 治療の状況について述べる。DMD 治療については，2016 年にヒトジストロフィンのメッセンジャーリボ核酸前駆体を標的とし，エクソン 51 スキッピングを誘導してジストロフィン蛋白質を回復させることを目的としたアンチセンスオリゴヌクレオチド（ASO）である EXONDYS 51®が，臨床的ベネフィットを予測する可能性がかなり高いと考えられる代替評価項目（ジストロフィン濃度）を用いて，迅速承認制度の下で FDA により承認された。エクソン 53 スキッピングに関する他の ASO も同様の理由で迅速承認を受けた。市販後にこれらの薬剤の臨床

的有用性を確認することは，米国 FDA の継続的な承認の条件である。EMA は，重篤または生命を脅かす疾患の治療を目的とした医薬品についても迅速承認制度を有しているが（条件付き製造販売承認と呼ばれる），DMD の治療を目的とした医薬品の有効性を確立するための適切なバイオマーカーは存在せず，承認には臨床症状に関する評価項目に関するデータが必要であるとしている[6]。

Suvodirsen について

Suvodirsen はエクソン 51 スキッピングを誘発させるアンチセンス・オリゴヌクレオチドである。以下に記載する DYSTANCE 51 試験で承認申請を計画していたが，第 I 相試験の結果からジストロフィン発現増加のエビデンスは認められなかったことから，DYSTANCE 51 試験の中止を含め，Suvodirsen の開発中止の決定に至っている。しかし DYSTANCE 51 試験はベイズ流の解析による有効性または安全性の欠如による試験の中止，症例数の変更，投与群の中止，用量の併合，外部対照の借用など様々な要素でのアダプテーションが検討されており，かつ CID パイロットプログラムに採択されていることから，本報告書で言及するに値すると考え以下に記載する。

DYSTANCE 51 試験

DYSTANCE 51 試験は歩行可能な DMD 患者を対象としたランダム化プラセボ対照二重盲検並行群間試験である。48 週間の二重盲検期，及び 48 週間の非盲検期からなる試験であった。150 例までの DMD 患者を 1:1:1 でプラセボ，suvodirsen 低用量，高用量に割り付けられた。主要評価項目は規制当局間で異なっている。FDA に対して用いられた主要評価項目は筋生検で得られたジストロフィンのベースラインからの変化量であり，これは迅速承認のために用いられる代替評価項目である。一方で PMDA，及び EMA に対しては 48 週時の North star ambulatory assessment（NSAA）のベースラインからの変化量であった。NSAA は DMD を対象とした行動評価表であり，17 項目に 0~2 点（0 が Unable to achieve indepenvdently，2 が Normal に対応する）をつけることにより，最高 34 点のスコアで評価される。症例数 150 例は 48 週時点の NASS で群間差を 3，標準偏差を 4.5，脱落率を 10%と想定し，有意水準を両側 0.05 とすると，検出力を 88%得られるという観点で設定された。なおこの場合，ジストロフィンでの検出力は群間差を 4.0%，標準偏差を 3.0%，脱落率を 10%と想定し，有意水準を両側 0.05 とすると 99%以上であった[7]。

中間解析

本試験では複数の中間解析が計画された（図 3-6-1-2-1）。ジストロフィンに関する中間解析は 2 回計画され，1 回目の中間解析（D1）は最初の 30 例の患者が投与後 12 週間後のジストロフィンに関するデータが得られた時点で計画された。また 2 回目の中間解析（D2）は次の 40 例の患者（すなわち 31 例目から 70 例目までの）患者が投与後 22 週間後のジストロフィンに関するデータが得られた時点で計画された。この中間解析では①FDA での迅速承認を目的とした有効性の評価，②

低用量群の中止，③低用量群，及び高用量群の併合（NSAA についてのみ）が検討される予定であった。一方 NSAA に関する中間解析は 70, 90, 110, 130 例で計画されており（N1~N4），中間解析の結果から最終解析の成功確率を算出し，症例登録中止の検討がされていた。中間解析では外部情報も用いられているが，その詳細は後述する。

ジストロフィンを測定するための筋生検は各患者 2 回のみ実施する計画であった。これは筋生検の侵襲性が高いことが原因と考えられる。すべての被験者はベースラインで筋生検が行われるが，試験薬投与開始後のジストロフィン観測時点は被験者が登録されたタイミングによって変化する。1 回目のジストロフィンに関する中間解析（D1）に含まれる最初の 30 例は 12 週目に，2 回目のジストロフィンに関する中間解析（D2）に含まれる次の 40 例は 22 週目に，それ以降の 80 例は 46 週目にジストロフィンを測定する。なお NSAA は 12 週ごとに評価されるが，筋生検の実施が NSAA の評価に影響を及ぼす可能性があるので，NSAA の評価時点（24 週及び 48 週）の 2 週間前に実施している。

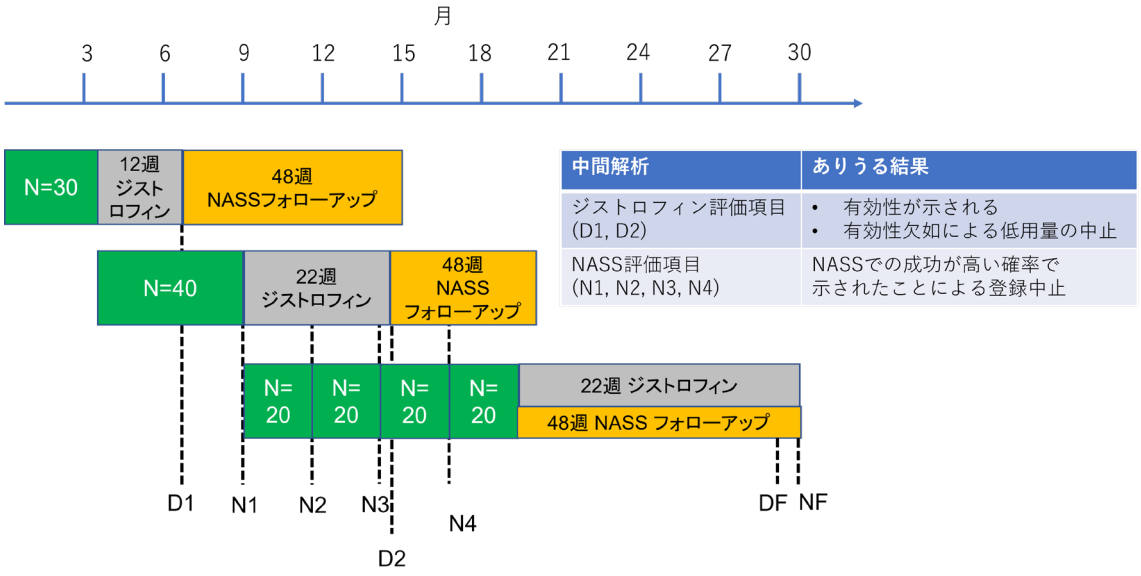


図 3-6-1-2-1 DYSTANCE 51 試験デザイン
Lake (2021) [5]をもとに作成

<u>ジストロフィンに関する解析計画</u>

ジストロフィンに関する解析計画の詳細を述べる。ジストロフィンは以下に示すベイズ流線形回帰モデルに基づき解析を行う計画であった。

$$D_{ij} = \tau_j + \vartheta_{t(i),j} + \epsilon_{ij}$$
$$\epsilon_{ij} \sim N(0, \sigma^2_{t(i),j})$$

D_{ij}は患者iの時点j（1: 12 週, 2: 22 週, 3: 46 週）でのジストロフィン（単位%）の変化量，τ_jは時点jでのプラセボ群のジストロフィンの平均変化量，$t(i)$は患者iの治療群（0:プラセボ，

1:低用量，2:高用量），ϑ_{Tj}は時点jでの治療T（1:低用量，2:高用量）の効果，ϵ_{ij}は誤差項であり，各時点，及び治療群で異なる分布が設定されている。事前分布は無情報事前分布を仮定し，プラセボ群の各時点の平均変化量τ_j，及びϑ_{Tj}については

$$\tau_j \sim N(0,10^2); j = 1,2,3$$

$$\vartheta_{Tj} \sim N(0,10^2);\ T = 1,2, j = 1,2,3$$

と設定された。τ_j及びϑ_{Tj}の標準偏差に関しては過去に DMD を対象に承認された薬剤の臨床試験におけるジストロフィンのベースラインからの変化量の平均値が 0.28~5.9%であったことから，その最大値の約 2 倍である 10%を用い，無情報事前分布に近づけるようにしたことが紹介されている。実際にこういった事前分布を用いることで，τ_jの事前分布の 95%信用区間は-19.6%~+19.6%となり，過去の試験で観測された値を含んでいると考えることができる。各時点での標準偏差σ_{Tj}の事前分布は

$$\sigma_{Tj} \sim Unif(0,\ 10); T = 0,1,2,\ j = 1,2,3$$

であり，これは過去の臨床試験のジストロフィンのベースラインからの変化量の標準偏差は 0.40~4.5 であったこと，0 から 10 までの一様分布を用いることでこれらの値が十分に含まれていることから設定された。

加えてジストロフィンについては suvodirsen の第 I 相長期投与試験のジストロフィンに関する結果の情報が借用された。第 I 相長期投与試験では第 I 相単回投与試験で 12 週間の追跡調査期間を終了した患者 34 例を対象に，suvodirsen を 2 週間に 1 回投与した。主要目的は安全性評価であったが，ジストロフィンに関するデータも 12 週時または 22 週時に収集されていた。一回目の中間解析時には，第 I 相長期投与試験の高用量群で 12 週時，及び 22 週時の結果がそれぞれ 10 例，及び 15 例，低用量群で 22 週時の結果が 9 例利用可能であった。しかしながら第 I 相長期投与試験で 12 週時のジストロフィンの結果が得られている高用量群では，DYSTANCE 51 試験の高用量群よりも少ない用量が用いられていた。そのため有効性の評価には，第 I 相長期投与試験のデータは考慮されなかったが，用量の選択（前述の②低用量群の中止）においては加味された。

ジストロフィンに関する治療効果への比較は以下のように実施される計画であった。まず suvodirsen で 2 用量群あることへの対応としては Gatekeeping 法を用い，高用量で有意な結果の時にのみ，低用量での有効性を評価する手順で実施された。また 2 回の中間解析，及び最終解析，計 3 回の解析が実施される点に関しては Bonferroni 法を用い，片側有意水準 0.025/3 回 $= 0.00833$ が各解析時点の有意水準として用いられる。これをベイズ流の評価に反映して 1 回目の中間解析では$\Pr(\vartheta_{2,1} > 0|d) > 0.991667$，2 回目の中間解析では$\Pr(\vartheta_{2,2} > 0|d) > 0.991667$，最終解析の時点では$\Pr(\vartheta_{2,3} > 0|d) > 0.991667$であった場合に有効性があると考える。

中間解析では低用量群の中止が評価される。中間解析時に suvodirsen の高用量群で効果があり，かつ低用量群での効果が高用量の半分以下である可能性が高い場合に中止される。具体的には以下の場合に低用量群が中止する。

$$\Pr(\vartheta_{2,j} > 0\ \&\ \vartheta_{1,j} < 0.5\vartheta_{2,j}|d) > 0.95 \text{ for } j = 1,2$$

1 回目の中間解析では 12 週時，及び 22 週時のジストロフィンに関するデータから低用量群の中止が検討される。12 週時に関しては DYSTANCE 51 試験 12 週時データ，及び第 I 相長期投与試験のデータを利用して低用量群の中止が判断される。一方，DYSTANCE 51 試験の 22 週時のジストロフィンに関するデータは，1 回目の中間解析時にはまだ十分に取得できていないため，プラセボ群では 12 週と 22 週でジストロフィンの値は同じと想定(τ_1=τ_2)し，第 I 相長期投与試験の 22 週データと DYSTANCE 51 試験のプラセボ群 12 週時データで比較する。なおこの解析は低用量群の中止についてのみ実施され，有効中止には用いられない。2 回目の中間解析では DYSTANCE 51 試験の 22 週時のジストロフィンデータが取得できているので，そのデータを低用量群中止に関する無益性評価に用い，中止基準は同じである。なお，中間解析の結果，低用量群を中止した場合，既に低用量群に割り付けられた患者は高用量群へ移行するが，NSAA を含む臨床評価項目の解析には含めない。

NSAA に関する解析計画

NSAA に関する解析計画の詳細を述べる。NSAA はジストロフィンの解析とは独立で実施され，繰り返し測定値に対するベイズ流疾患進行モデルを用いてプラセボとの比較を行う計画であった。プラセボは後述する外部試験のデータも用い，ベイズ流メタ・アナリシスの方法を用いて補強する。データ間の異質性はベイズ流に評価し，結果が類似している場合はより多くの情報を借用する。また前述した低用量群の中止がなかった場合は，高用量群と低用量群は併合して解析する計画であった。

NSAA の解析には，以下の繰り返し測定値に対する疾患進行モデルを用いた。

$$Y_{ij} = \gamma_i + \exp\left(\theta_{t(i)} + \eta_i + \boldsymbol{\alpha}' \boldsymbol{X}_i + \delta_{s(i)}\right) \sum_{k=1:j} \beta_k + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma^2_{t(i),s(i)})$$

Y_{ij}は患者iの時点j（0: ベースライン， 1: 12 週， 2: 24 週， 3: 36 週， 4: 48 週）での NSAA スコア，γ_iは患者iのランダム効果，$t(i)$は患者iの治療群（0:プラセボ， 1:低用量, 2:高用量），$\theta_{t(i)}$はプラセボと比較した治療効果，η_iは各患者での推移の違いに関するランダム効果，$\boldsymbol{X_i}$はC個の共変量，$\delta_{s(i)}$は背景情報の違いで説明できないデータソース間のランダム効果，β_kはプラセボ群での各時点間の平均変化に該当する（なお$\beta_0 = 0$とする）。また$s(i)$はそのデータのソースとなった試験を意味し，0 は DYSTANCE 51 試験，$1 \sim S$ は後述するヒストリカルデータである。ϵ_{ij}は治療群，データソースごとに異なる標準偏差を持つ独立な正規分布を想定している。このモデルに基づくと，各時点間の NASS の変化がβ_kによってモデル化されている。例えばプラセボ群のある被験者iのベースライン時点，及び各時点の NASS はそれぞれ

$$Y_{i0} = \gamma_i + \epsilon_{i0}$$

$$Y_{i1} = \gamma_i + \exp\left(\theta_{t(i)} + \eta_i + \boldsymbol{\alpha}' \boldsymbol{X}_i + \delta_{s(i)}\right) \beta_1 + \epsilon_{i1}$$

$$Y_{i2} = \gamma_i + \exp\left(\theta_{t(i)} + \eta_i + \boldsymbol{\alpha}'\boldsymbol{X}_i + \delta_{s(i)}\right)(\beta_1 + \beta_2) + \epsilon_{i2}$$
$$Y_{i3} = \gamma_i + \exp\left(\theta_{t(i)} + \eta_i + \boldsymbol{\alpha}'\boldsymbol{X}_i + \delta_{s(i)}\right)(\beta_1 + \beta_2 + \beta_3) + \epsilon_{i3}$$
$$Y_{i4} = \gamma_i + \exp\left(\theta_{t(i)} + \eta_i + \boldsymbol{\alpha}'\boldsymbol{X}_i + \delta_{s(i)}\right)(\beta_1 + \beta_2 + \beta_3 + \beta_4) + \epsilon_{i4}$$

で与えられる。$\exp\left(\theta_{t(i)} + \eta_i + \boldsymbol{\alpha}'\boldsymbol{X}_i + \delta_{s(i)}\right)\beta_k$が時点$k-1$から時点$k$までの NASS の変化量の期待値を表していることがわかる。患者iがプラセボ群，低用量群に割り付けられた場合の 12 週目の NSAA のベースラインからの変化量の期待値は，

$$E(Y_{i1} - Y_{i0}|t(i) = 0) = \exp\left(\eta_i + \boldsymbol{\alpha}'\boldsymbol{X}_i + \delta_{s(i)}\right)\beta_1$$
$$E(Y_{i1} - Y_{i0}|t(i) = 1) = \exp\left(\theta_1 + \eta_i + \boldsymbol{\alpha}'\boldsymbol{X}_i + \delta_{s(i)}\right)\beta_1$$

で与えられ，

$$\frac{E(Y_{i1} - Y_{i0}|t(i) = 1)}{E(Y_{i1} - Y_{i0}|t(i) = 0)} = \exp(\theta_1)$$

となるため，治療効果$\theta_{t(i)}$は疾患の進行の比（Disease Rate Ratio; DRR）を表していることになる。なお，これは 24 週以降も同様であり，各週で治療効果が一定の疾患進行の比を表していると仮定している。

疾患進行モデルについてベイズ流に推定を行うために，利用可能な NASS に関する情報を用いた事前分布を設定する。ベースラインの NSAA に関しては以下の階層分布を用いてモデル化を行う。

$$\gamma_i \sim N(\mu_{\gamma,s(i)}, \sigma_\gamma^2); i = 1, \cdots, n$$
$$\mu_{\gamma,s} \sim N(20, 10^2); s = 0, \cdots, S$$
$$\sigma_\gamma \sim Unif(0, 20)$$

また，平均的な共変量をもつプラセボ群患者における 12 週ごとの進行の変化率は 0 以下の切断正規分布

$$\beta_k \sim N_{(-\infty,0]}(0, 2^2); j = 1,2,3,4$$

と設定している。これは DMD 患者において NASS が経時的に減少すると仮定していることになる。各データソース，各治療群の標準誤差$\sigma_{T,s}$は

$$\sigma_{T,s} \sim Unif(0, 10);\ T = 0,1,2; s = 0, \cdots, S$$

治療Tの効果に関しては，治療によって疾患が進行・減少いずれにも同じ重みを置いた分布，

$$\exp(\theta_T) \sim Unif(0, 2);\ T = 1,2$$

を事前分布として設定している。

各患者での推移の違いに関するランダム効果η_iは，以下に示す平均 1，標準偏差に一様分布の超事前分布を用いた階層ガンマ分布によってモデル化された。

$$\exp(\eta_i) \sim \text{Gamma}\left(1/\sigma_\eta^2, 1/\sigma_\eta^2\right) \text{ for } i = 1, \cdots, n$$
$$\sigma_\eta \sim Unif(0, 2)$$

また共変量の効果に関しては，C個の共変量に対してそれぞれの独立同分布の以下の事前分布を仮定している。

$$\alpha_c \sim N(0, 10^2); c = 1, \cdots, C$$

さらに今回複数のデータソースを用いて解析を実施しており，そのデータソース間の違いを表すδ_sに関しては

$$\exp(\delta_s) \sim \text{Gamma}\left(1/\sigma_\delta^2, 1/\sigma_\delta^2\right) \text{ for } s = 1, \cdots, S$$

$$\sigma_\delta \sim Unif(0, 1)$$

を事前分布として設定している。σ_δが小さい場合，より多くヒストリカルデータを借用することが可能である。

中間解析は 70, 90, 110, 130 例時点で実施する。事前にシミュレーションの結果に基づき決定した治療効果 DRR_N（70, 90, 110, 130 例時点でそれぞれ 0.50, 0.525, 0.55, 0.575）を満たす確率が高い場合（90%超）はそれ以降の症例登録を中止する。なお DRR_N は中間解析時点の症例数で，有効性が言えるために必要な治療効果の事後平均である。

$$\Pr(\exp(\theta_T) < DRR_N | y) > 0.9$$

最終解析では治療群間の比較として DRR が 1 未満である事後確率を評価する。$\Pr(\exp(\theta_2) < 1|y) > 0.975$の場合に NSAA に対する有効性が示せたこととなる。前述の通り，最終解析にあたり，低用量が中止していない場合には$\theta_1 = \theta_2$を仮定し，低用量群と高用量群を併合した集団で評価を実施する。また低用量群が中止している場合には，高用量群のみを評価することとなる。

最後に NSAA に関するヒストリカルデータの利用に関して述べる。DYSTANCE 51 試験では DMD 患者を対象とした 3 試験のプラセボデータが用いられている。これらはすべて他社で実施された臨床試験であり，tadalafil の第 III 相試験（Tadalafil DMD），ataluren の第 III 相試験（ACT DMD），及び domagrozumab の第 II 相試験（B5161002）である。Tadalafil と ataluren では Critical Path Institute's（C-Path's） Duchenne Regulatory Science Consortium（D-RSC）の枠組みも利用してデータを取得している。ただし，D-RSC の持つデータベースに含まれていないデータセットへのアクセスは容易ではなく，Domagrozumab に関しては試験を実施したファイザー社と協議したことが論文中に記載されている。ヒストリカルデータの利用可能性は，過去に実施された臨床試験と DYSTANCE 51 試験での交換可能性に依存する。交換可能性を検討するにあたり①客観的かつ一貫した盲検下の評価，②一貫した適格・除外基準，③一貫した標準・背景治療，④対照群，及び予後因子の可能性があるベースライン因子の違いに対する調整の事前規定，に関して検討がされた。上記の 3 試験において，DYSTANCE 51 試験の主要な適格基準がいずれのデータセットでも設定されていること，ベースラインと 48 週時の NASS は一貫していたことから，交換可能性を支持するものと考えられた。

表 3-6-1-1 NSAA に関するヒストリカルプラセボデータ

データセット	例数	NSAA の ベースライン値 (SD)	NSAA のベースラインから の変化量 (SD)	残差誤差
Tadalafil DMD	90	22.5 (6.0)	-4.3 (3.7)	2.4
ACT-DMD	76	24.0 (5.8)	-3.5 (3.4)	2.2
B5161002	26	20.6 (5.9)	-4.6 (5.8)	3.2

Lake (2021) [5]をもとに作成

シミュレーションによる動作特性評価

Lake (2021) [5]では上記の試験デザインに関するシミュレーション結果も報告している。シミュレーションに用いたパラメータとしては，被験者の登録速度，ジストロフィンのベースライン平均及び測定誤差，高用量と低用量での有効性の比，ジストロフィンに対して最大の治療効果を発揮するまでの期間，第 I 相長期投与試験と DYSTANCE 51 試験の患者でのジストロフィンの値の差異，NASS の平均及び標準偏差の想定値（ベースライン及び投与後）が設定されている。この条件下で動作特性（試験の成功確率，早期の試験成功確率，平均症例数，平均試験期間，低用量群の脱落確率，治療効果の推定値の中央値，情報借用された有効サンプルサイズ）が検討されている。詳細は Lake (2021) [5]を参考にしていただきたい。

まとめ

本節では希少疾病において有効性または安全性の欠如による試験の中止，投与群の中止，用量の併合，外部対照の借用などの要素をすべて組み込んだ臨床試験について紹介をした。こういった複数のモチベーションに関しては，本来であれば探索的な臨床試験を別途実施したのちに，検証的な試験へと進む場合が多かった。しかし希少疾病では臨床試験に登録可能な症例数が限られているため，複数の試験を実施するのは難しい場合がある。そういった場合に，様々な観点を組み合わせた複雑な臨床試験デザインが考えられ，効率的な医薬品開発につながる可能性がある。FDA のガイダンスでも強調されているように，CID を用いた試験の動作特性はシミュレーションにより十分に評価する必要がある。

参考文献

[1] Food and Drug Administration. "Federal register notice, Complex Innovative Designs Pilot Meeting Program." (2018). https://www.federalregister.gov/documents/2018/08/30/2018-18801/complex-innovative-designs-pilot-meeting-program

[2] Food and Drug Administration. "Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products- Guidance for Industry " (2020).

[3] Food and Drug Administration. Statement by former FDA Commissioner Scott Gottlieb on FDA's

new steps to modernize drug development, improve efficiency, and promote innovation of targeted therapies, https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-fdas-new-steps-modernize-drug-development-improve

[4] Price, Dionne, and John Scott. "The US Food and Drug Administration's Complex Innovative Trial Design Pilot Meeting Program: Progress to date." Clinical Trials 18.6 (2021): 706-710.

[5] Lake, Stephen L., et al. "Bayesian adaptive design for clinical trials in Duchenne muscular dystrophy." Statistics in Medicine 40.19 (2021): 4167-4184.

[6] European Medicines Agency. "Guideline on the clinical investigation of medicinal products for the treatment of Duchenne and Becker muscular dystrophy." (2015).

[7] ClinicalTrials.gov. " Efficacy and Safety Study of WVE-210201 (Suvodirsen) With Open-label Extension in Ambulatory Patients With Duchenne Muscular Dystrophy (DYSTANCE 51)." NCT03907072 https://clinicaltrials.gov/ct2/show/NCT03907072

3.6.2 N-of-1 デザイン

N-of-1 デザインとは，1 人の患者に対し時期を違えてランダムに複数の治療法を割り付け，その効果を検討する試験デザインである（以降，N-of-1 デザインを使用した臨床試験を N-of-1 試験と呼ぶ）。このため，その症例の特徴，疾患の状態，併用治療，評価内容等を加味した，同一個体内（同一条件内）での治療効果の推定・治療選択が可能になることが期待される。一方で，被験者個人の状況や評価方法のばらつきが個々の試験結果に影響するため，単独の試験結果の一般化が難しい場合が考えられる。この点については後述の通り，複数の N-of-1 試験をメタ・アナリシスにより統合することによって，集団での治療効果を検討できる可能性がある。例えば，Amitriptyline の若年性特発性関節炎による痛みをアウトカムとする臨床試験では，ベイズ流メタ・アナリシスにより 6 つの N-of-1 試験を統合し，その治療法が評価されている[1]。また，複数の N-of-1 試験をもとに，小規模集団における被験者間の効果のばらつきを検討できるため，Proof of concept の検討にも活用できる可能性がある[2]。ただし，N-of-1 デザインの利用にあたっては，慢性疾患であることや，治療開始に伴い効果が迅速に現れ，終了により迅速に消失する等，実装可能性が高い状況において選択すべきであろう[3]。

N-of-1 デザインの方法論，及び解析手法等に関する近年の発展はガイドとしてまとめられている[4, 5]。以降，本ガイド，及び関連論文をもとに，N-of-1 デザインの概略，及び解析手法を紹介する。

3.6.2.1 デザイン概略

N-of-1 デザインは，1 人の患者を対象とした，患者内ランダム化二重盲検多重クロスオーバーにより複数の治療を比較する試験デザインである[4]。1 人の患者を対象とする N-of-1 試験は，患者個人における効果（Individual Treatment Effect，以下 ITE）を推定し，個別治療を最適化することに焦点を置く。個人における治療の意思決定のためのエビデンスを評価する研究デザインの階層において，N-of-1 デザインは最上位に位置づけられている[6]。

典型的な N-of-1 試験の例を図 3-6-2-1-1 に示す。この例は，合計 6 つの治療ピリオドを設定し，治療 A と治療 B を比較することを想定している。6 つの治療ピリオドは，2 種類の治療を連続して実施する治療ピリオドのまとまり（サイクル）に分かれる。図 3-6-2-1-1 の例では，1 サイクルが 2 つの治療ピリオドで構成され，その合計サイクル数は 3 となる。1 人の被験者について，各サイクルに A→B，または B→A の順に治療を実施する順序をランダムに割り付け，それぞれの治療ピリオドでアウトカムを測定する。

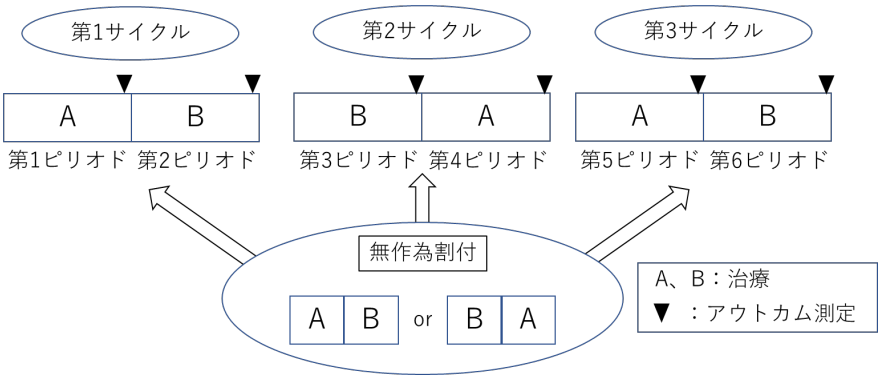


図 3-6-2-1-1 N-of-1 デザインの例

このように，N-of-1 デザインでは，1 人の患者がランダムな順番で複数の治療を受け，さらにそれを理想的には少なくとも 3 回以上繰り返し実施する。この点が，各患者が一つの治療のみを受ける典型的な並行群間比較試験との大きな違いと言える。

3.6.2.2　N-of-1 試験の統合

個々の N-of-1 試験の目的は，ITE を推定して患者個人における治療のエビデンスを示すことであるが，類似した患者を対象に同じ治療を評価した複数の N-of-1 試験を統合し，母集団の治療効果を推定する手法が提案されている。このような手法は，Meta-analysis of N-of-1 trials，Aggregation of N-of-1 Trials，Combined N-of-1 Trials，Series of N-of-1 Trials，Serial N-of-1 Trials 等と表現される。すなわち，N-of-1 試験の統合によって，臨床試験への症例登録が困難な希少疾患領域であっても，母集団での治療効果を推定できる利点があると考えられている[7]。また，特定の条件のもとで，N-of-1 試験の統合により，並行群間比較試験に比して症例数が減少することが期待される（3.6.2.3, 及び補遺で詳述）。

N-of-1 試験の統合の実装にあたっては，各患者の結果を統合することを前提としたプロトコルに基づいて行われることが多い。参加する患者全員に適用する同一の試験計画及び統合解析計画をプロトコルに予め規定し，N-of-1 試験を前向きに実施して結果を統合した事例は，多々見受けられる。また，統合に用いる解析手法のアイデアは，1 人の患者の結果を 1 つの試験とみなしたメタ・アナリシスである。同一の選択基準，介入方法，アウトカムの測定方法等を規定したプロトコルを使用することにより，類似患者を対象とした同じデザインの試験結果を統合する理想的なメタ・アナリシスが可能となる。なお，特に希少疾患領域の場合は，1 人の患者を対象とした過去の N-of-1 試験の成績が豊富に公表される状況にないため，上記のように複数人を対象に N-of-1 試験を新たに実施し，その結果を統合することが盛んに行われているものと推察される。

希少疾患領域の医薬品開発の観点から，N-of-1 試験の統合の適用場面として，以下の点が挙げられる。なお，3.～5.は一般的なクロスオーバーデザインの適用可能な条件である。

- 適用場面
  1. 希少疾患領域の患者母集団が非常に少なく，通常の並行群間比較試験や 2×2 クロスオーバー試験が不可能な場合に，少人数のデータにより母集団の治療効果を推定する[4]。
  2. 1 人 1 人の被験者について，治療期間やアウトカム測定を増やすことが許容可能である。
  3. 症状が安定した慢性的な疾患である[6]。
  4. 迅速に効果が現れる，かつ投与終了により迅速に効果が消失する治療である[3, 6]。
  5. ウォッシュアウト期間を短く設定できる（薬剤の半減期が短い）[6]。

また，N-of-1 試験の統合の長所，短所，限界として，以下の点が挙げられる。

- 長所
  1. 並行群間比較試験に比べ，症例数の削減が見込まれる[4]。
  2. 各被験者において，試験治療のアウトカムと対照治療のアウトカムのマッチングデータを複数（サイクルの数）獲得し，交絡を最小化し，個人差を考慮して分散を減少させる[8]。
  3. 脱落が起こった場合，脱落前のマッチングされたアウトカムのデータを最大限利用できる。
  4. プラセボ対照の場合であっても，実薬投与が確約された倫理面へ配慮したデザインにより，症例登録の促進が期待される[8]。

- 短所，限界
  1. 被験者の試験期間が長くなることで，脱落しやすくなる[5, 8]。
  2. 試験期間が長くなるほど，アウトカムの欠測が増加する。
  3. 症例数が少ないため，一般化可能性の主張が難しい[4]。

3.6.2.3 解析手法

N-of-1 試験の統合の解析手法は複数提案されている。以降，Summary Fixed and Random Effects model，Mixed model，Mixed Effects Model of Difference，Bayesian model を紹介する。各モデルは，連続型のアウトカムに適用することを想定する。なお，各モデル式の表記を原著から変えていることに注意されたい。

i. Summary Fixed and Random Effects model

Summary Fixed Effects Model, 及び Summary Random Model は，各 N-of-1 試験の要約された治療効果データを用いて，複数の N-of-1 試験の結果を統合するメタ・アナリシスモデルである。第 i 番目の N-of-1 試験（第 i 被験者）で推定された治療効果（試験治療－対照治療）をy_iとする。Summary Fixed Effects model は，y_iが全体の平均治療効果の真値αを中心にばらつくと仮定する方法であり，次のように定式化される[4]。

$$y_i = \alpha + \varepsilon_i\ ;\ \varepsilon_i \sim N(0, \sigma_i^2)$$

σ_i^2は，y_iに仮定した正規分布の分散である。試験ごとの繰り返し測定数または治療ピリオド数が少ない場合，σ_i^2の推定精度は十分には見込めないが，各試験が類似したデザインであれば試験共通の等分散性（$\sigma_i^2 = \sigma^2$）を仮定した解析が行われる。

一方，Summary Random Effects model では，y_iが試験（被験者）特異的な治療効果α_iを中心にばらつくと仮定する。α_iは全体の治療効果α_0を平均，τ^2を分散とする正規分布に従うランダム効果であり，次のように定式化される[4]。

$$y_i = \alpha_i + \varepsilon_i\ ;\ \varepsilon_i \sim \mathrm{N}\left(0, \sigma_i^2\right); \alpha_i \sim N(\alpha_0, \tau^2)$$

また，等価な別表現としてモデルを次のように表せる。

$$y_i = \alpha_0 + \alpha_i + \varepsilon_i\ ;\ \varepsilon_i \sim N\left(0, \sigma_i^2\right);\ \alpha_i \sim N(0, \tau^2)$$

ここで，σ_i^2は治療効果の試験内分散（被験者内分散），τ^2は治療効果の試験間分散（被験者間分散）を表現している。被験者間分散τ^2を推定するためには十分な数の N-of-1 試験が必要となるが，そうではない場合はτ^2の推定が困難になる可能性があるため，Summary Fixed Effects model が好ましいとされる[9]。

以上の方法は，その名に”Summary”とあるように，過去に実施された N-of-1 試験（被験者）ごとに要約された治療効果の推定値のデータを使用することを想定している。しかしながら，希少疾患領域においては，過去に実施された臨床試験の結果は十分に蓄積されておらず，上記のモデルの適用は現実的に難しい可能性がある。よって，希少疾患領域の医薬品開発の観点では，次に紹介する Mixed Model がより現実的であると考えられる。

ii. Mixed Model

Mixed Model は，被験者レベルのデータを用いて複数の N-of-1 試験の結果を統合するメタ・アナリシスモデルである。希少疾患領域においては，過去に実施された臨床試験が少ないため，治療やアウトカムの測定等が同一のデザインの N-of-1 試験を，前向きに複数試験実施し（複数人で実施し），得られた各被験者のデータに対してモデルを適用することが想

定される。Schmid and Duan (2014) [10]による Mixed Model を以下に示す。

$$y_{imjk\ell} = \alpha_i + \beta_\ell + \gamma_k + \delta_{j(k)} + \varepsilon_{m\left(j(k(i))\right)}$$

$$\alpha_i \sim N(\alpha_0, \sigma_\alpha^2);\ \gamma_k \sim N\left(0, \sigma_\gamma^2\right);\ \delta_{j(k)} \sim N\left(0, \sigma_\delta^2\right);\ \varepsilon_{m\left(j(k(i))\right)} \sim N(0, \sigma_\varepsilon^2)$$

上記モデルは，各ピリオドでアウトカムを複数回測定する経時測定デザインを想定している。$y_{imjk\ell}$は，第i被験者が第kサイクル内の第jピリオドにおいて治療ℓを受け，同ピリオドで第m回目に測定されたアウトカムである。α_iは試験（被験者）効果，β_ℓは治療効果，γ_kはサイクルの効果，$\delta_{j(k)}$は第kサイクル内のピリオドjの効果，$\varepsilon_{m\left(j(k(i))\right)}$はランダム誤差である。$\beta_\ell$は固定効果であり，他の項は上式の平均・分散の正規分布に従うと仮定したランダム効果である。

上記の Mixed Model を拡張し，ピリオドに線形的なトレンドを仮定すること，また，治療の持ち越し効果を入れることも可能である。さらには，誤差項εの時間を通じた相関，ピリオドに対する非線形の時期トレンド，季節効果，被験者と他の因子の交互作用等を考慮したモデルへの拡張が可能である[10]。

iii. Mixed Effects Model of Difference

各ピリオドのアウトカムに対する Mixed Model ではなく，治療間のアウトカムの差に対するよりシンプルなモデルとして，Mixed Effects Model of Difference が提案されている。Mixed Effects Model of Difference では，各サイクルにおける 2 治療のアウトカムの差のデータを目的変数として，次式のような混合効果モデルを仮定する[11-13]。

$$y_{ik} = \alpha + \gamma_k + \alpha_i + \varepsilon_{ik}$$
$$\alpha_i \sim N(0, \tau^2);\ \varepsilon_{ik} \sim N(0, 2\sigma^2)$$

ここで，y_{ik}は第i被験者の第kサイクルにおけるアウトカムの治療間差であり，被験者間，及びサイクル間で独立と仮定している。切片αは全被験者共通の母平均，γ_kは第 k サイクルの固定効果，α_iは第i被験者の治療効果であり，平均0と分散τ^2の正規分布を仮定したランダム効果である。なお，各ピリオドのアウトカムのランダム誤差が正規分布に従うとし，その平均，及び分散をそれぞれ 0，及びσ^2とすると，サイクル内の治療間差y_{ik}のランダム誤差ε_{ik}の分散は，$2\sigma^2$と表現される。

iv. ベイズモデル

上述のモデルは未知パラメータを定数として扱う頻度流の方法である。これに対し，未知パラメータに事前分布を導入し，事後分布により推論を行うベイズモデルが提案されている[9, 10, 14-16]。

Zucker et al. (2006) [15]による階層ベイズモデルは，第i被験者の第jピリオドのアウトカムY_{ij}に対して次のモデルを仮定する。

$$Y_{ij} \sim N(\alpha_i + \beta_i X_{ij}, \sigma_i^2)$$

ここで，X_{ij}は治療を表すダミー変数，α_iは第i被験者の切片，β_iは第i被験者の治療効果，σ_i^2は被験者ごとに仮定したy_{ij}の分散である。α_i，及びβ_iはランダム効果として扱い，階層構造をなすように，$\alpha_i \sim N(\alpha_0, \tau_\alpha^2)$，$\beta_i \sim N(\beta_0, \tau_\beta^2)$とそれぞれ仮定し，さらに，$\alpha_0$，$\beta_0$，$\tau_\alpha^2$，$\tau_\beta^2$，$\sigma_i^2$の各パラメータに対して事前分布を仮定する。データにより事前分布を更新して事後分布を獲得し，集団の治療効果β_0，及び個人の治療効果β_iの推論を行う。また，事後確率を用いて，「治療効果が臨床的に意義のある閾値を上回る確率」を直接的に評価することが行われている。

階層ベイズモデルを用いてN-of-1試験を統合するアプローチの利点として，集団，及び個人それぞれにおける治療効果の推定が可能であり，これらの推定に事前情報を利用できることが挙げられる[14]。また，本アプローチにより推定した個人の治療効果の事後平均は，他の被験者の情報を利用して調整（縮小推定）される[14]。後述の実例では，集団だけでなく個人の治療効果の推定も重視したものが多く，このような特徴から階層ベイズモデルがN-of-1試験の統合に頻繁に用いられているものと思われる。

階層ベイズモデルを用いてN-of-1試験を統合した報告は，これまでに多数なされている[1, 17-25]。そのうち，希少疾患領域の参考として，非ジストロフィー性ミオトニー症候群の患者を対象としたN-of-1試験統合の実例を挙げる[24, 25]。同研究は，38名の患者を対象に単一のプロトコルに基づくN-of-1試験を前向きに実施し，その結果を階層ベイズモデルにより統合することで，筋肉のこわばりの重症度スコアに対するMexiletineの有効性を評価した。同研究では，階層ベイズモデルによる各被験者，及び集団の治療効果の推論がMexiletineの有効性を支持していること，また，慢性的な希少疾患の介入評価におけるN-of-1試験の実施可能性が主張されている。なお，ベイズモデルでは，パラメータの事前分布の設定方法がしばしば議論になる。同研究は，過去のクロスオーバー試験の結果に基づく事前分布を主解析に使用し，専門家の意見を反映した事前分布，及び無情報事前分布をそれぞれ仮定した感度解析を実施しているため，ベイズモデルを実装する際の留意点の参考にされたい。なお，本研究は，各サイクルを完了する度に被験者単位で中間解析を実施し，治療効果の閾値に対する事後確率を評価することで，早期有効中止または無効中止を判定している。指定難病のような希少疾患領域でN-of-1試験を検討する際は，このように倫理面に配慮した柔軟なベイズ流デザインは参考になると考えられる。デザイン，及び解析手法の詳細については，同文献を参照いただきたい。

3.6.2.4　症例数設定

N-of-1試験の統合においては，必要症例数として，実施試験数（症例数）だけではなく，各試験（被験者）におけるサイクル数の選び方が重要となる。N-of-1試験の統合に必要な症

例数設定方法については，補遺でいくつか紹介する。

3.6.2.5　典型的な臨床試験デザインとの比較

検出力，及び選択バイアスの脆弱性の観点から，N-of-1 試験の統合と，典型的な並行群間比較試験，2×2 クロスオーバー試験の性能比較を行った知見を紹介する。Blackston et al. (2019) [26]は，「3.6.2.3 ii Mixed model」に述べたモデルを仮定したシミュレーションを行い，3 種類のデザインを同人数で実施する場合，検出力は N-of-1 試験で最も高く，αエラーは N-of-1 試験で名目有意水準（0.05）に最も近いことを示した。一方，3.6.2.2 にも述べた通り，N-of-1 試験の短所として，症例数が少ない場合に一般化可能性の主張が困難になり得ることが挙げられる。この点に関して Blackston et al. (2019) [26]は，選択バイアスに対する N-of-1 デザインの脆弱性を定量的に評価した。例えば，「標的母集団では治療無効」が真であるが，治療が有効な異なる母集団の被験者が試験に誤って混入することで，選択バイアスが発生する状況をシミュレートしている。標的母集団，及び非標的母集団それぞれのサンプリング確率を変化させたシミュレーションの結果，N-of-1 デザインでは，他の典型的なデザインに比してαエラーが顕著に増大した[26]。原因として，N-of-1 デザインでは同一被験者から繰り返しデータを取得することから，他のデザインよりもデータ数が増えて標準誤差が小さくなるために，αエラーが増大した可能性が考えられる。これは治療効果の推論を行う上で，「バイアスが大きく，かつ，精度が高い」という最も好ましくない状況と言えよう。

以上の知見から Blackston et al. (2019) [26]は，少人数で高い検出力が見込める N-of-1 デザインは，質の高いエビデンスを創出すると考えられる一方で，標的母集団からのサンプリングには特に配慮が必要であると強調している。確かに，市販後に薬剤を使用する母集団に対して，N-of-1 試験に組み入れる被験者が少ない場合ほど，選択バイアスが入らないように選択・除外基準を十分に検討した上で，統合結果の一般化にも慎重にならなければならない。これに対し，母集団の大部分が標本となるような極めて患者が少ない希少疾患領域では，選択バイアスは問題になりにくく，N-of-1 試験は高い検出力を以って治療のエビデンスを示す選択肢になりうると考えられる。

参考文献

[1] Huber, Adam M., et al. “Amitriptyline to relieve pain in juvenile idiopathic arthritis: a pilot study using Bayesian metaanalysis of multiple N-of-1 clinical trials.” The Journal of rheumatology 34.5 (2007): 1125-1132.

[2] Hilgers, Ralf-Dieter, et al. “Lessons learned from IdeAl—33 recommendations from the IdeAl-net about design and analysis of small population clinical trials.” Orphanet journal of rare diseases 13.1 (2018): 1-17.

[3] Margolis, Amanda, and Christopher Giuliano. “Making the switch: From case studies to N-of-1 trials.” Epilepsy & behavior reports 12 (2019): 100336.

[4] Nikles, Jane, and Geoffrey Mitchell, eds. The essential guide to N-of-1 trials in health. New York, NY, USA:: Springer, 2015.

[5] Kravitz, R. L. “and the DEcIDE Methods Center N-of-1 Guidance Panel (Duan N, Eslick I, Gabler NB, Kaplan HC, Kravitz RL, Larson EB, Pace WD, Schmid CH, Sim I, Vohra S).” Design and implementation of N-of-1 trials: a user’s guide. AHRQ Publication 13 (2016): 14.

[6] Guyatt G, Jaeschke R, McGinn T. Therapy and Validity: N-of-1 Randomized Controlled Trials. Users’ Guides to the Medical Literature: A manual for Evidence-Based Clinical Practice. Chicago, IL: American Medical Association, 2002: 275-290.

[7] Facey, Karen, et al. “Generating health technology assessment evidence for rare diseases.” International journal of technology assessment in health care 30.4 (2014): 416-422.

[8] Gupta S, Faughnan ME, Tomlinson GA, Bayoumi AM. A framework for applying unfamiliar trial designs in studies of rare 6diseases. J Clin Epidemiol. 2011 Oct;64(10):1085-94.

[9] Zucker, Deborah R., Robin Ruthazer, and Christopher H. Schmid. “Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations.” Journal of clinical epidemiology 63.12 (2010): 1312-1323.

[10] Schmid, C. H., and N. Duan. “Chapter 4: The DEcIDE Methods Centre N-of-1 guidance panel statistical design and analytic consideration for N-of-1 trials.” Kravitz RL, Duan N, The DEcIDE Methods Centre N-of-1 Guidance Panel (Duan N, Eslick L, Gabler NB, Kaplan HC, Kravitz RL, Larson EB, Pace WD, Schmid CH, Sim I, Vohra S)(eds) Design and implementation of N-of-1 Trials: a user’s guide. AHRQ Publication 13 (2014): 14.

[11] Chen, Xinlin, and Pingyan Chen. “A comparison of four methods for the analysis of N-of-1 trials.” PloS one 9.2 (2014): e87752.

[12] Araujo, Artur, Steven Julious, and Stephen Senn. “Understanding variation in sets of N-of-1 trials.” PloS one 11.12 (2016): e0167167.

[13] Senn, Stephen. “Sample size considerations for n-of-1 trials.” Statistical methods in medical research 28.2 (2019): 372-383.

[14] Zucker, D. R., et al. “Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment.” Journal of clinical epidemiology 50.4 (1997): 401-410.

[15] Zucker, Deborah R., et al. “Lessons learned combining N-of-1 trials to assess fibromyalgia therapies.” The Journal of rheumatology 33.10 (2006): 2069-2077.

[16] Duan, Naihua, Richard L. Kravitz, and Christopher H. Schmid. “Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research.” Journal of clinical epidemiology 66.8 (2013): S21-S28.

[17] Coxeter, P. D., et al. "Valerian does not appear to reduce symptoms for patients with chronic insomnia in general practice using a series of randomised n-of-1 trials." Complementary therapies

in medicine 11.4 (2003): 215-222.

[18] Nathan, P. C., et al. “A pilot study of ondansetron plus metopimazine vs. ondansetron monotherapy in children receiving highly emetogenic chemotherapy: a Bayesian randomized serial N-of-1 trials design.” Supportive care in cancer 14.3 (2006): 268-276.

[19] Sung, Lillian, et al. “Serial controlled N-of-1 trials of topical vitamin E as prophylaxis for chemotherapy-induced oral mucositis in paediatric patients.” European Journal of Cancer 43.8 (2007): 1269-1275.

[20] Yelland, M. J., et al. “Celecoxib compared with sustained-release paracetamol for osteoarthritis: a series of n-of-1 trials.” Rheumatology 46.1 (2007): 135-140.

[21] Yelland, Michael J., et al. “N-of-1 randomized trials to assess the efficacy of gabapentin for chronic neuropathic pain.” Pain Medicine 10.4 (2009): 754-761.

[22] Nikles, Jane, et al. “Do pilocarpine drops help dry mouth in palliative care patients: a protocol for an aggregated series of n-of-1 trials.” BMC palliative care 12.1 (2013): 1-7.

[23] Nikles, Catherine J., et al. “Aggregated n-of-1 trials of central nervous system stimulants versus placebo for paediatric traumatic brain injury–a pilot study.” Trials 15.1 (2014): 1-11.

[24] Stunnenberg, Bas C., et al. “Combined N-of-1 trials to investigate mexiletine in non-dystrophic myotonia using a Bayesian approach; study rationale and protocol.” BMC neurology 15.1 (2015): 1-10.

[25] Stunnenberg, Bas C., et al. “Effect of mexiletine on muscle stiffness in patients with nondystrophic myotonia evaluated using aggregated N-of-1 trials.” Jama 320.22 (2018): 2344-2353.

[26] Blackston JW, Chapple AG, McGree JM, McDonald S, Nikles J. Comparison of Aggregated N-of-1 Trials with Parallel and Crossover Randomized Controlled Trials Using Simulation Studies. Healthcare (Basel). 2019 Nov 6;7(4):137.

3.6.3　最近の議論

本節では，特に近年提案されている複数の手法を紹介する。実際の新薬開発で活用された経験はなかったものの，どのような状況下であれば適用が検討できるか，また，そのメリット・デメリットなどについても考察する。

3.6.3.1　Complete n-of-1 デザイン

3.6.3.1.1　　デザイン概略

クロスオーバーデザインに関する最近の議論として，Complete n-of-1 デザインが提案されている[1-4]。Complete n-of-1 デザインは，順序群の設定を工夫した多重クロスオーバーデザインであり，複数人の被験者を対象に実施する。各ピリオドで実施する治療の全ての組合せを含むように順序群を設定し，被験者をランダムに割り付ける。その具体的な例として，試験治療（T：Test）と対照治療（R：Reference）の有効性を比較する場合にあって，ピリオドがそれぞれ 2，3，4 期の Complete n-of-1 デザインの順序群を表 3-6-3-1-1 に示す。表の通り，例えばピリオド数が 4 の場合，$2^4 = 16$通りの順序群に被験者をランダムに割り付け，各ピリオドにおいてアウトカムを測定することとなる。

表 3-6-3-1-1 Complete n-of-1 デザイン　順序群の例

順序群	ピリオド1	ピリオド2	ピリオド3	ピリオド4
1	R	R	R	R
2	R	T	R	R
3	T	T	R	R
4	T	R	R	R
5	R	R	T	R
6	R	T	T	T
7	T	R	T	R
8	T	T	T	T
9	R	R	R	T
10	R	R	T	T
11	R	T	R	T
12	R	T	T	R
13	T	R	R	T
14	T	R	T	T
15	T	T	R	T
16	T	T	T	R

Complete n-of-1 デザインは多重クロスオーバーデザインの 1 種であるため，3.6.2.2 に述べた n-of-1 試験の特徴のうち，典型的なクロスオーバーデザインと同様の特徴をもつと考えられる。

また，Chow(2020)[4]は，先発バイオ医薬品に対する後発品（バイオシミラー）の互換性（Interchangeability）を示すための生物学的同等性試験において，Complete n-of-1 デザイン

の使用を提案している。バイオシミラーの互換性に関しては 3.6.3.1.3 に詳述する。

3.6.3.1.2 Complete n-of-1 デザインの定式化及び推定量の導出

Complete n-of-1 デザインは，多重クロスオーバー試験として一般化した枠組みにより，アウトカムを定式化する。K順序群，Jピリオドの$K \times J$クロスオーバー試験において，次のような線形混合効果モデルを仮定する[3] [4]。

$$y_{ijk} = \mu + G_k + S_{ik} + P_j + D_{d(j,k)} + C_{d(j-1,k)} + e_{ijk}$$
$$i = 1,2,\dots n_k;\ j = 1,2,\dots J;\ k = 1,2,\dots K;\ d = T\ or\ R$$

ここで，y_{ijk}は，第k順序群の第i被験者が，第jピリオドにおいて治療dを受けた時のアウトカムである。μは全体平均，G_kは第k順序群の効果，S_{ik}は第k順序群の被験者iのランダム効果（平均値 0, 分散σ_s^2），P_jは第jピリオドの効果，$D_{d(j,k)}$は第k順序群の第jピリオドにおける治療dの効果，$C_{d(j-1,k)}$は第k順序群の第$(j-1)$ピリオドに治療dを受けた時の持ち越し効果，e_{ijk} はランダム誤差（平均 0, 分散σ_e^2）である。S_{ik}とe_{ijk}は互いに独立と仮定する。

続いて，$K = 16$順序群，$J = 4$ピリオドの16×4 Complete n-of-1 デザインを例に, Chow (2018)[2]の方法による治療効果$D_T - D_R$の推定量の導出を解説する。Chow (2018)[2]の方法は, 上記モデルの(k, j)の組合せ（表 3-6-3-1-1 の 64 個のセル）におけるアウトカムの標本平均の線形結合により, $D_T - D_R$の推定量を構成する。各セルの標本平均の重みを表すベクトル$\boldsymbol{\beta}$, セルの標本平均のベクトル$\overline{\boldsymbol{Y}}$を以下の通り定義する。

$\boldsymbol{\beta}' = \{\beta_{jk}\}$

$=(\beta_{11}, \beta_{21},\ \beta_{31}, \beta_{41},\ \beta_{12}, \beta_{22}, \beta_{32}, \beta_{42}, \dots, \beta_{1\,15}, \beta_{2\,15}, \beta_{3\,15}, \beta_{4\,15},\ \beta_{1\,16}, \beta_{2\,16}, \beta_{3\,16}, \beta_{4\,16})$

$\overline{\boldsymbol{Y}}' = (\bar{Y}_{\cdot 11}, \bar{Y}_{\cdot 21}, \bar{Y}_{\cdot 31}, \bar{Y}_{\cdot 41}, \bar{Y}_{\cdot 12}, \bar{Y}_{\cdot 22}, \bar{Y}_{\cdot 32}, \bar{Y}_{\cdot 42}, \dots, \bar{Y}_{\cdot 1\,15}, \bar{Y}_{\cdot 2\,15}, \bar{Y}_{\cdot 3\,15}, \bar{Y}_{\cdot 4\,15}, \bar{Y}_{\cdot 1\,16}, \bar{Y}_{\cdot 2\,16}, \bar{Y}_{\cdot 3\,16}, \bar{Y}_{\cdot 4\,16})$

$$\bar{Y}_{\cdot jk} = \frac{\sum_{i=1}^{n_k} Y_{ijk}}{n_k}$$

治療効果$D_T - D_R$の推定量$\widetilde{D}$を次のように表し, 重み$\boldsymbol{\beta}$を求めて推定量を構成することが目標となる。

$$\widetilde{D} = \boldsymbol{\beta}'\overline{\boldsymbol{Y}}$$

また，線形混合効果モデルの固定効果のパラメータベクトル$\boldsymbol{P}$，治療効果を推定するための対比ベクトル$\boldsymbol{L}$を

$\boldsymbol{P}' = (\mu, G_1, G_2, G_3, G_4, G_5, G_6, G_7, G_8, G_9, G_{10}, G_{11}, G_{12}, G_{13}, G_{14}, G_{15}, G_{16}, P_1, P_2, P_3, P_4, D_T, D_R, C_T, C_R)$

$\boldsymbol{L}' =(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,-1,0,0\ )$

と定義すると，$\widetilde{D}$が不偏推定量であるためには，次式が成立しなければならない。

$$E[\widetilde{D}] = E[\boldsymbol{\beta}'\overline{\boldsymbol{Y}}] = \boldsymbol{L}'\boldsymbol{P}$$

一方で，16×4 Complete n-of-1 デザインにおけるデザイン行列$\boldsymbol{X}$, 及び$\boldsymbol{P}$を用いて，各セル

のアウトカムの期待値$\boldsymbol{\omega}$を表す表現を考える。

$$\boldsymbol{\omega}' = \{\omega_{jk}\}$$
$$= (\omega_{11}, \omega_{21}, \omega_{31}, \omega_{41}, \omega_{12}, \omega_{22}, \omega_{32}, \omega_{42}, \dots, \omega_{1\,15}, \omega_{2\,15}, \omega_{3\,15}, \omega_{4\,15}, \omega_{1\,16}, \omega_{2\,16}, \omega_{3\,16}, \omega_{4\,16})$$

とすると，

$$\boldsymbol{E}[\overline{\boldsymbol{Y}}] = \boldsymbol{X}'\boldsymbol{P} = \begin{bmatrix} 1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&1&0&0 \\ 1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&1 \\ & & & & & & & & \vdots & & & & & & & & & & & \\ 1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&1&0&1&0&1&0 \\ 1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&1&1&0 \end{bmatrix} \begin{bmatrix} \mu \\ G_1 \\ \vdots \\ G_{16} \\ P_1 \\ \vdots \\ P_4 \\ D_T \\ D_R \\ C_T \\ C_R \end{bmatrix}$$

$$= \begin{bmatrix} \mu + G_1 + P_1 + D_R \\ \mu + G_1 + P_2 + D_R + C_R \\ \vdots \\ \mu + G_{16} + P_3 + D_T + C_T \\ \mu + G_{16} + P_4 + D_R + C_T \end{bmatrix} = \begin{bmatrix} \omega_{11} \\ \omega_{21} \\ \vdots \\ \omega_{3\,16} \\ \omega_{4\,16} \end{bmatrix} = \boldsymbol{\omega}$$

以上より，$E[\boldsymbol{\beta}'\overline{\boldsymbol{Y}}] = \boldsymbol{\beta}'\boldsymbol{\omega} = \boldsymbol{L}'\boldsymbol{P} \rightarrow \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{P} = \boldsymbol{L}'\boldsymbol{P} \rightarrow \boldsymbol{\beta}'\boldsymbol{X}' = \boldsymbol{L}' \rightarrow \boldsymbol{L} = \boldsymbol{X}\boldsymbol{\beta}$ となるため，この方程式を$\boldsymbol{\beta}$について解き，$\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{L}$を計算することにより，推定量の係数$\boldsymbol{\beta}$が得られる。本タスクフォースにおいて，実際に計算した結果を表 3-6-3-1-2 に示す。なお，便宜的に整数で示すため，$\boldsymbol{\beta}$に 264 を掛けた値を示している。

表 3-6-3-1-2　16×4 Complete n-of-1 デザインの治療効果の推定量に使用する係数（本タスクフォースにて導出）

264×β（治療は表3-6-3-1-1に対応）				
順序群	ピリオド1	ピリオド2	ピリオド3	ピリオド4
1	3	-1	-1	-1
2	-5	15	-1	-9
3	11	15	-9	-17
4	19	-1	-9	-9
5	-5	-9	15	-1
6	-19	1	9	9
7	11	-9	7	-9
8	-3	1	1	1
9	-3	-7	-7	17
10	-11	-15	9	17
11	-11	9	-7	9
12	-13	7	15	-9
13	13	-7	-15	9
14	5	-15	1	9
15	5	9	-15	1
16	3	7	7	-17

表 3-6-3-1-2 に基づき，治療効果の推定量$\widetilde{D}$とその期待値，及び分散は次のように導出される。

$$\widetilde{D} = \boldsymbol{\beta}'\bar{\boldsymbol{Y}} = 1/264 \times (3\bar{Y}_{.11} - \bar{Y}_{.21} - \bar{Y}_{.31} - \bar{Y}_{.41} - 5\bar{Y}_{.12} + 15\bar{Y}_{.22} - \bar{Y}_{.32} - 9\bar{Y}_{.42} + \cdots + 5\bar{Y}_{.1\,15} + 9\bar{Y}_{.2\,15} - 15\bar{Y}_{.3\,15} + \bar{Y}_{.4\,15} + 3\bar{Y}_{.1\,16} + 7\bar{Y}_{.2\,16} + 7\bar{Y}_{.3\,16} - 17\bar{Y}_{.4\,16})$$

$$E[\widetilde{D}] = D_T - D_R, \quad V[\widetilde{D}] = \frac{\sigma_e^2}{11n}$$

ここで，nは順序群当たりの症例数である。

3.6.3.1.3 FDA における互換性の見解と Switching デザイン

互換性については FDA が明確な考えを示しており，バイオシミラーであることに加え，「どのような患者においても先発バイオ医薬品と同じ臨床効果を示すこと」と「反復投与される製品の場合，バイオシミラーと先発バイオ医薬品の切り替え時の安全性上のリスク，及び有効性減弱のリスクが，切り替えずに先発バイオ医薬品を使用する場合のリスクよりも大きくないこと」を以って，当該バイオシミラーに互換性があると判断される[5]。互換性のあるバイオシミラーとして FDA に承認されれば，処方する医師の介入を経ずに先発品との切り替えが可能となる[5]。また，FDA (2019)[5]は，切り替え実施群（switching arm）と非実施群（non-switching arm）で互換性を評価する Switching デザインを想定し，さらに，前者の群に通常少なくとも 3 回の切り替えを設けることを推奨するが，スポンサーが別のアプローチを提案する場合はこの限りではない。米国においては 2021 年 7 月に，FDA が互換性のある糖尿病治療薬としてのインスリン製剤バイオシミラー（商品名 Semglee，Mylan Pharmaceuticals 社）を初めて承認したところであり[6]，互換性を評価した実例は少ない状況である。なお，Semglee の互換性は，Switching デザインではなく，先発品に対する有効性の非劣性を検証した 2 本の並行群間比較試験（INSTRIDE1，INSTRIDE2）により評価されている（ClinicalTrials.gov Identifier: NCT02227862，NCT02227875）。各試験ともに約 500 名の規模で実施されたことから，対象となる糖尿病患者の症例登録は比較的容易であったものと推察される。同規模の並行群間比較試験の実施が困難な希少疾患領域においては，Complete n-of-1 のような多重クロスオーバーデザインが有用となり得る。

3.6.3.1.4 Complete n-of-1 デザインと Switching デザインの比較

上述のとおり，FDA は切り替えを 3 回，すなわち表 3-6-3-1-1 の 11 及び 1 の順序群のみを設定する Switching デザイン（以降，(RTRT，RRRR)デザイン）を推奨している。(RTRT，RRRR)デザインの順序群を表 3-6-3-1-4-1 に示す。

表 3-6-3-1-4-1 (RTRT，RRRR)デザイン　順序群

順序群	ピリオド1	ピリオド2	ピリオド3	ピリオド4
1	R	R	R	R
2	R	T	R	T

また，3.6.3.1.2 と同様の考え方により，(RTRT，RRRR)デザインの治療効果の推定量$\widetilde{D}$に使用する係数を表 3-6-3-1-4-2 に示す。

表 3-6-3-1-4-2　(RTRT，RRRR)デザインの治療効果の推定量に使用する係数[2]

$2\times\beta$（治療は3-6-3-1-4-1に対応）				
順序群	ピリオド1	ピリオド2	ピリオド3	ピリオド4
1	2	-1	0	-1
2	-2	1	0	1

したがって，(RTRT，RRRR)デザインの$\widetilde{D}$及び $V\left[\widetilde{D}\right]$は次式のとおり表せる。

$$\widetilde{D} = \boldsymbol{\beta}'\overline{\boldsymbol{Y}} = 1/2 \times (2\overline{Y}_{.11} - \overline{Y}_{.21} - \overline{Y}_{.41} - 2\overline{Y}_{.12} + \overline{Y}_{.22} + \overline{Y}_{.42})$$

$$V\left[\widetilde{D}\right] = \frac{3\sigma_e^2}{n}$$

Chow(2020) [4]によれば，(RTRT, RRRR)デザインに比べ，16×4 Complete n-of-1 デザインでは標準誤差の減少による症例数削減が見込め，効率が良いとされる。Chow (2020)[4]は，16×4 Complete n-of-1 デザインの(RTRT，RRRR)デザインに対する相対効率の例として，全体で 48 例（Complete n-of-1：3 例/順序群，（RTRT，RRRR）：24 例/順序群）の試験を実施する場合，2 つのデザインの$V\left[\widetilde{D}\right]$の比が 24.24%となることを示し，$16 \times 4$ Complete n-of-1 デザインの効率が良いと主張している。

希少疾患領域の医薬品開発において，「患者が少ないながらも，如何にしてエビデンスを示すか」が大きな課題であることは言うまでもない。Chow(2020) [4]の提案は，バイオシミラーの互換性評価という限定的な状況を想定してはいるものの，希少疾患領域の臨床試験の必要症例数の削減に貢献しうる重要な知見であると考えられる。

3.6.3.1.5　Complete n-of-1 デザインの解析手法

Complete n-of-1 デザインは，先発バイオ医薬品に対するバイオシミラーの互換性評価のために提案されたデザインであり，互換性を示すためには同等性検定が行われる。16×4 Complete n-of-1 デザインにおいて，3.6.3.1.2 で導出した$\widetilde{D}$に基づき，事前に設定した同等性マージンθについて，同等性仮説（$H_0: |D_T - D_R| > \theta \; versus \; H_1: |D_T - D_R| \leq \theta$）を検証する仮説検定を構成できる。検定統計量$T_D$が以下の条件を満たす場合に，$H_0$を棄却する[3, 4]。

$$T_D = \frac{\widetilde{D} - \theta}{\hat{\sigma}_e^2\sqrt{\frac{1}{11n}}} > t\left[\frac{\alpha}{2}, 16n - 5\right]$$

また，$D_T - D_R$の（$1-\alpha$）%信頼区間は次式により与えられる[3, 4]。

$$\tilde{D} \pm t\left[\frac{\alpha}{2}, 16n-5\right]\hat{\sigma}_e^2\sqrt{\frac{1}{11n}}$$

一般化した$K \times J$クロスオーバーデザインについて，推定量$\tilde{D}$に基づく同等性検定における症例数設定の方法[3, 4]は補遺で紹介する。

まとめ

3.6.3.1 節では，Complete n-of-1 デザインの概略及び解析手法を示し，バイオシミラーの互換性評価における FDA 推奨のデザインに対する効率性の議論を紹介した。Complete n-of-1 デザインは比較的新しく提案されたこともあり，本タスクフォースは事例調査を行ったものの，医薬品開発の承認申請に使用された実例は確認できなかった。また，上述の通り，Complete n-of-1 デザインは，必要症例数を抑え，希少疾患領域のバイオシミラーの互換性評価を効率的に実施することを目的に提案された。

希少疾患におけるバイオシミラーの互換性評価に限らず，新薬開発への Complete n-of-1 デザインの適用を検討することは有益かもしれない。まず，Complete n-of-1 デザインを適用可能な疾患・治療の前提条件として，クロスオーバーデザインを適用できる一般的特徴（3.6.2.2 参照）を持つことが挙げられる。この条件をクリアしたとして，新薬開発を目的とした比較試験に Complete n-of-1 デザインを採用する場合，同一治療を長期間継続する群が存在することに留意しなければならない。例えば，上述した16×4 Complete n-of-1 デザインでは，TTTT，及び RRRR のように同一治療を 4 ピリオドにわたって継続する順序群を設定する。重篤な症状が慢性的に現れる希少疾患領域において，このような群設定が倫理的な観点から問題とならないかを十分に考慮しなければならない。また，実施可能性の観点からは，被験者集積が困難になる懸念や，脱落が多発する懸念がある。

さらに，効率性の観点からは，バイオシミラーの互換性（同等性）評価に推奨される(RTRT, RRRR)デザインよりむしろ，新薬の優越性・非劣性検証を想定した並行群間比較試験や 2×2 クロスオーバー試験を基準として，Complete n-of-1 の効率を比較することが求められる。すなわち，被験者集積が容易であれば本来選択するはずの典型的デザインに比べ，それが実施不可能な場合，Complete n-of-1 を採用して優越性・非劣性検定を行うことで症例数がどれほど減少するかを吟味し，希少疾患の新薬開発における有用性を検討することも一案と考える。

参考文献

[1] Chow, Shein-Chung, Fuyu Song, and Can Cui. “On hybrid parallel–crossover designs for assessing drug interchangeability of biosimilar products.” Journal of biopharmaceutical statistics 27.2 (2017):

265-271.

[2] Chow, Shein-Chung, ed. Encyclopedia of Biopharmaceutical Statistics-Four Volume Set. 564-71. CRC Press, 2018.

[3] Chow, Shein-Chung, and Yu-Wei Chang. “Statistical considerations for rare diseases drug development.” Journal of biopharmaceutical statistics 29.5 (2019): 874-886.

[4] Chow, Shein-Chung. Innovative Methods for Rare Disease Drug Development. 185-204. CRC Press, 2020.

[5]Food and Drug Administration. “Considerations in demonstrating interchangeability with a reference product-guidance for industry.” Food and Drug Administration (2019).

[6]Food and Drug Administration. “FDA Approves irst Interchangeable Biosimilar Inslin Product for Treatment of Diabetes.” FDA NEWS RELEASE(2021).

Available URL at: https://www.fda.gov/news-events/press-announcements/fda-approves-first-interchangeable-biosimilar-insulin-product-treatment-diabetes.

3.6.3.2　2 段階アダプティブデザイン

本節では, Chow and Huang [1]で提案されている 2 段階アダプティブデザインについて紹介する。近年, 論文で発表された手法であり, 実際の治験への適用可能性は未知である。新薬の承認を得るためには, 医薬品の安全性と有効性に関する実質的な証拠を提供する必要がある。実際には, プラセボ対照試験を実施し, 以下の仮説を検証するのが典型的なアプローチである。

H_0: ineffectiveness versus H_a: effectiveness　(1)

ineffectiveness※1 という帰無仮説の棄却は, effectiveness※2 という対立仮説を支持していることになる。しかし, effectiveness の支持は, effectiveness の証明を意味するものではないことに注意すべきである。Chow and Huang [1]では, 仮説(1)が次のようになるべきだと主張している。

H_0: ineffectiveness versus H_a: not ineffectiveness　(2)

(1)と(2)のH_aからわかるように effectiveness と not ineffectiveness※3 の概念は同じではなく, not ineffectiveness は effectiveness を意味しない。つまり, 概念的には not ineffectiveness には inconclusiveness※4 と effectiveness の部分が含まれる（図 3-6-3-2-2-1 を参照）。図 3-6-3-2-2-1 に示すように,（未知の）真の治療効果をθとし,（θ_L, θ_U）をθの対応する（$1-\alpha$）×100%信頼区間とすると, 仮説(1）は次の(3)のように書き換えられる。つまり, 帰無仮説は「$\theta \leq \theta$の信頼区間の下限」で, 対立仮説は「$\theta >$$\theta$の信頼区間の上限」で表記できる。

H_0: $\theta \leq \theta_L$ versus H_a: $\theta > \theta_U$ (3)

一方, 仮説(2)は次のような非劣性に関する典型的な片側検定となる。

H_0: $\theta \leq \theta_L$ versus H_a: $\theta > \theta_L$ (4)

したがって, 帰無仮説の棄却は非劣性の結論につながり, 同等性（inconclusiveness, すなわち $\theta_L < \theta \leq \theta_U$）と優越性（すなわち effectiveness）からなる。与えられた症例数に対して, 帰無仮説が棄却された場合にのみ, その医薬品が not ineffectiveness であることを証明することができる。本当に効果があることを証明するためには, 結論が出ない可能性をなくすために当該治療の inconclusiveness を棄却する必要がある。

※1 ineffeciveness：有効性は認められない
※2 effectiveness：臨床的に意義がある有効性
※3 not ineffectiveness：有効性が認められないことはない
※4 incoclusiveness：臨床的な意義について結論できない

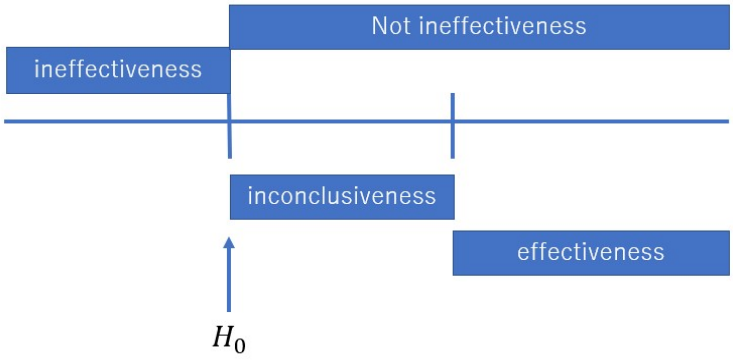


図 3-6-3-2-2-1 not ineffectiveness, inconclusiveness と effectiveness の関係

3.6.3.2.1 2 段階アダプティブデザインの手順と解析手法

希少疾患では，症例登録できる患者が少ないため，規制当局の審査や承認において有効性を検証するための基準の定め方は，希少疾患治療薬開発における最も大きな課題である。そういった問題に対処するために，希少疾患治療薬の開発の Stage1 で not ineffectiveness であることを証明し，Stage2 で effectiveness であることを証明する，2 段階アダプティブデザインを説明する。2 段階アダプティブデザインの概要は以下の通りである。

Stage1: 過去/パイロット試験または文献調査に基づき分散の推定値σ^2，θの$(1-\alpha)\times 100\%$信頼区間を構築する。次に，Stage1 で得られたn_1名の被験者をもとに，事前規定された有意水準α_1で非劣性の仮説（例えば，not ineffectiveness）に対して検定を行う。具体的には，以下の検定統計量を作成する。

$$T_1 = T(X_{n_1}) = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i}$$

ここで$X_{n_1} = \{X_{11}, X_{12}, \cdots, X_{1n_1}\}$は，$n_1$名の被験者のアウトカムである。$\Phi(\cdot)$を標準正規分布の累積分布関数とし，対応する p 値をP_1とすると，$1-\Phi\left(\frac{T_1-\theta_L}{\sigma/\sqrt{n_1}}\right)$と同等かそれ以下であり，

$P_1 \le \alpha_1$ 又は $T_1 \ge \theta_L + z_{\alpha_1}\frac{\sigma}{\sqrt{n_1}}$の場合は ineffectiveness の帰無仮説を棄却する。Ineffectiveness の帰無仮説を棄却できなかった場合は無効として試験を中止する。その他の場合は試験を継続し，次の Stage に進む。

Stage2: Stage2 でさらにn_2名の被験者を追加登録する。この Stage では，試験途中の治療の有効性を証明するための望ましい統計的な保証（例えば検出力 80%）を達成するために，症例数の再設定を行うことができる。Stage2 では，予め指定された有意水準α_2で，inconclusiveness の領域内の確率が許容範囲内であることを保証するために統計的な検定を行う。具体的には追加の症例数n_2のアウトカムを$X_{n_2} = \{X_{21}, X_{22}, \cdots, X_{2n_1}\}$，検定統計量を

$T_2 = T(X_{n_2}) = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$, 下記のように推定された inconclusiveness の確率を$\widehat{P_I}$とする。

$$\widehat{P_I} = \left(\Phi\left(\frac{\theta_U - T_2}{\sigma/\sqrt{n_1}} + z_{\alpha_1}\right) - \Phi\left(\frac{\theta_L - T_2}{\sigma/\sqrt{n_1}} + z_{\alpha_1}\right)\right) \Big/ \left(1 - \Phi\left(\frac{\theta_L - T_2}{\sigma/\sqrt{n_1}} + z_{\alpha_1}\right)\right)$$

B 個のブートストラップ標本, 検定統計量, inconclusiveness の確率をそれぞれ $\{X_{n_2}^1, X_{n_2}^2, \cdots, X_{n_2}^B\}$, $\{T_2^1, T_2^2, \cdots, T_2^B\}$, $\{\hat{P}_I^1, \hat{P}_I^2, \cdots, \hat{P}_I^B\}$とし$P_2 = \frac{1}{B}\sum_{b=1}^{B} I\{\hat{P}_I^b < \hat{P}_I\}$とする。このとき, $P_2 \leq \alpha_2$の場合 effectiveness を主張することができる。

2 段階アダプティブデザインの下では, 第 1 種の過誤確率は$(\alpha_1, \beta_1, \alpha_2)$の関数で示される。ただし, β_1は Stage1 での無効中止の基準とする。そのため, α_1及びβ_1を適切に選択すれば, not ineffectiveness であることを証明するために必要な症例数を少なくすることができる。ただし, $(\alpha_1, \beta_1, \alpha_2)$の選択は, 試験実施前に治験実施計画書に事前規定することが重要であり, 試験後の選択は推奨されない。2 段階アダプティブデザインの考え方は, Stage1 で事前規定された有意水準α_1で not ineffectiveness を示し, その後事前規定された有意水準α_2で effectiveness であることを結論づけるというものである。2 段階アダプティブデザインでは, method of individual p-values（MIP）, method of sum of p-values（MSP）, method of product of p-values（MPP）などの p 値に基づく検定統計量を用いると, 第 1 種の過誤確率は $(\alpha_1, \beta_1, \alpha_2)$ の関数となる （Chang, 2007）[2]。例えば, MIP について考えると, 第k stage の検定統計量は次のように与えられる。

$$P_k, k = 1, \cdots, K$$

$K = 2$の 2 段階アダプティブデザインでは, $\alpha = \alpha_1 + \alpha_2(\beta_1 - \alpha_1)$となる。そのため, 試験継続判断の基準は下記となる。

有効性中止	if $P_k \leq \alpha_k$
無益性中止	if $P_k > \beta_k$
Continue with adaptations	if $\alpha_k < P_k \leq \beta_k$

$(\alpha_1, \beta_1, \alpha_2)$を適切に選択することで, 第 1 種の過誤確率を制御することができる。MIP（Chow and Chang, 2006 and 2008）[3] [4]に基づいた, α_1, α_2, β_1のいくつかの組み合わせを表 3-6-3-2-1-1 に示す。表からわかるように, α_1とβ_1をそれぞれ 0.005 と 0.30 とした場合, α_2は 0.1525 で与えられ, 5%の有意水準の下で第 1 種の過誤確率を制御することができる。

表 3-6-3-2-1-1: Stopping boundaries of α_2 with MIP for one sided α=0.05

β_1 / α_1	0.000	0.005	0.100	0.015
0.10	0.5000	0.4737	0.4444	0.4118
0.15	0.3333	0.3103	0.2857	0.2593
0.20	0.2500	0.2308	0.2105	0.1892
0.30	0.1667	0.1525	0.1379	0.1228
0.75	0.0667	0.0604	0.0541	0.0476

β_1/α_1	0.000	0.005	0.100	0.015
1.00	0.0500	0.0452	0.0404	0.0355

3.6.3.2.2 2 段階アダプティブデザインの数値例

Chow and Huang [1]が紹介する 2 段階アダプティブデザインの数値例を下記に示す。2 段階アダプティブデザインでは, Stage1 で not ineffectiveness を証明し, Stage2 で effectiveness であることを結論づけることができる。例としてライエル病の細胞治療について考える。この疾患の発症率は, 欧州では人口 100 万人あたり 2 人と推定されている（Miller et al.2018[5]）。この疾患は急性疾患で欧州での死亡率は約 22%である。有効性主要評価項目は, 7 日目の全癒であり, 現在の治療法の奏効率は$\theta_c = 0.5$である。治療費が高額であることから, 新しい細胞治療の効果を検証するために, 単群の第 I/II 相試験を計画する。新しい治療の奏効率を$\theta_t = 0.6$と仮定し, 有意水準を片側 5%, 検出力を 80%とする。このとき, 従来の方法による症例数は$\bar{\theta} = (\theta_t + \theta_c)/2 = 0.55$から$N_0 = \bar{\theta}(1-\bar{\theta})\left(z_\alpha + z_\beta\right)^2/(\theta_t - \theta_c)^2 = 153$となる。$N = 153$が与えられた下で, θ_tの$(1-\alpha)\times 100\%$信頼区間は, $(\theta_L, \theta_U) = (0.53, 0.67)$となる。

2 段階アダプティブデザインを考えると Stage1 では, $H_0: \theta_t \le \theta_L = 0.53$, $H_a: \theta_t > \theta_L$; Stage2 では, 対立仮説がH_a: $\theta_t > \theta_U$=0.67 となる。$(\alpha_1, \beta_1, \alpha_2)$の組み合わせが与えられたとき, 第 1 種の過誤確率αと事前規定した検出力$(1-\beta)$を維持するための Stage1 の症例数n_1と Stage2 の症例数n_2を計算する。Simon's two-stage optimal design と同様にθが与えられたときの帰無仮説を受容する確率を定義する。

$$\Pr(T_1 \le c_1|\theta, n_1) + \int_{T_1 > c_1}^{\infty} f_{T_1}(t|\theta, n_1) \Pr(T_2 \epsilon (c_{21}, c_{22})|T_1 = t, \theta, n_2) dt$$
$$= \Pr(T_1 \le c_1|\theta, n_1) + E_{T_1 > c_1}[\Pr(T_2 \epsilon (c_{21}, c_{22})|T_1, \theta, n_2)] \quad (5)$$

Stage1 で早期有効中止を行わないとすると, α_1 =0 となる。$c_1 = \theta_L + z_{\beta_1}\sqrt{\frac{\theta_L(1-\theta_L)}{n_1}}$, $c_{21} = -\infty$, $c_{22} = \theta_L + z_{\alpha_2}\sqrt{\frac{\theta_L(1-\theta_L)}{n_2}}$とすると, (5)の式は下記のようになる。

$$\Pr(T_1 \le c_1|\theta, n_1) + \Pr(T_1 > c_1|\theta, n_1) \Pr(T_2 \le c_{22}|\theta, n_2) \quad (6)$$

帰無仮説の下限は, $\theta = \theta_L$なので前述の$\alpha = \alpha_1 + \alpha_2(\beta_1 - \alpha_1)$の式において α_1=0 の場合を考えると$\alpha = \alpha_2\beta_1$なので $1-\beta_1+\beta_1(1-\alpha_2)=(1-\alpha)$となり, 第 1 種の過誤確率を制御できる。また, 対立仮説の下限を$\theta = \theta_L$としたとき(6)の式は, $\sigma_L = \sqrt{\theta_L(1-\theta_L)}$, $\sigma_U = \sqrt{\theta_U(1-\theta_U)}$なので, 以下のようになる。

$$\phi\left(\frac{\theta_L - \theta_U}{\sigma_U/\sqrt{n_1}} + z_{\beta_1}\frac{\sigma_L}{\sigma_U}\right) + \left(1 - \phi\left(\frac{\theta_L - \theta_U}{\sigma_U/\sqrt{n_1}} + z_{\beta_1}\frac{\sigma_L}{\sigma_U}\right)\right)\phi\left(\frac{\theta_L - \theta_U}{\sigma_U/\sqrt{n_2}} + z_{\alpha_1}\frac{\sigma_L}{\sigma_U}\right) \quad (7)$$

$n_2 = \lambda n_1$とすると総症例数は,$N = (1+\lambda)n_1$となる。ここでは, 対立仮説のもとで(7)の式がβより大きくならないn_1の最小値を求めたい。表 3-6-3-2-2-2-1 には$(\alpha_1, \beta_1, \alpha_2)$がそれぞれ与えられときの帰無仮説の下での期待総症例数ENと(λ, n_1, N)のいくつかの可能な組み合わせを表している。

表 3-6-3-2-2-2-1 から, $(\alpha_1, \beta_1, \alpha_2)$の異なる組み合わせにおいて，妥当な範囲から$\lambda$の値を選択すれば, 2 段階アダプティブデザインの症例数は, Simon's two-stage design における症例数よりは多くなるが，従来の症例数の設定よりも少ない総症例数を必要とし，十分に少ない Stage1 の症例数，及び帰無仮説の下で少ない期待総症例数となる。具体的に表 3-6-3-2-2-2-1 からλの値を 1 付近で選択すると，総症例数が従来の症例数の設定に比べて少なくなることがわかる。

表 3-6-3-2-2-2-1：検出力 80%を維持するための$(\lambda,\ n_1, N)$ の組み合わせ。θ_L =0.53, $\theta_U = 0.67, \alpha = 0.05,\ \beta$=0.2 とする。EN は帰無仮説の下での期待総症例とし，従来の方法で必要な症例数は$N_0 = 153$となる。また, Simon's two-stage optimal, minimax design それぞれにおける，第 1 段階の症例数は 24 例, 31 例，総症例数は 91 例, 78 例となる。

λ	0.33	0.5	1	2	3
$\alpha_1 = 0, \beta_1 = 0.15, \alpha_2$=0.33					
n_1	88	72	56	49	48
N	117	108	112	147	192
EN	113	103	104	132	170
$\alpha_1 = 0, \beta_1 = 0.2, \alpha_2$=0.25					
n_1	103	77	54	42	39
N	137	116	108	126	156
EN	130	108	97	109	133
$\alpha_1 = 0, \beta_1 = 0.25, \alpha_2$=0.20					
n_1	119	85	54	39	34
N	158	128	108	117	136
EN	148	117	94	98	110

2 段階アダプティブデザインは，Stage1 で事前規定した有意水準α_1で対象となる試験薬が not ineffectiveness であることを証明し，not ineffectiveness が証明された後，Stage2 では有意水準α_2の下で effectiveness であることを結論づけるというものである。提案された 2 段階アダプティブデザインは，適切な$\alpha_1, \alpha_2, \beta_1$, 及び Stage1 と Stage2 の症例数を設定することで第 1 種の過誤確率を制御した上で総症例数を従来の症例数よりも少なくすることができる。そのため，症例登録できる患者が少ない希少疾病医薬品開発に利用できる可能性がある。また, Stage1 の段階で試験を早期に中止することができるため，時間とリソースを節約できる可能性がある。

参考文献

[1] Shein-Chung Chow, and Zhipeng Huang. “Demonstrating effectiveness or demonstrating not ineffectiveness–A potential solution for rare disease drug product development?.” Journal of Biopharmaceutical Statistics 29.5 (2019): 897-907.

[2] Chang, Mark. "Adaptive design method based on sum of p‐values." Statistics in medicine 26.14 (2007): 2772-2784.

[3] Chow, Shein-Chung, and Mark Chang. Adaptive design methods in clinical trials. Chapman and Hall/CRC, 2006.

[4] Chow, Shein-Chung, and Mark Chang. "Adaptive design methods in clinical trials–a review." Orphanet journal of rare diseases 3.1 (2008): 1-13.

[5] Miller, Frank, et al. "Approaches to sample size calculation for clinical trials in rare diseases." Pharmaceutical statistics 17.3 (2018): 214-230.

3.6.3.3 Probability monitoring procedure

希少疾患の医薬品開発では, 症例登録可能な患者が少ないため検出力や推定精度に基づき試験の症例数を設定することが難しい場合がある。Huang 2019[1]は, そのような問題に対処するため, Probability monitoring procedure を提案している。本節では, この Probability monitoring procedure の説明を行う。Probability monitoring procedure とは, 実施可能性, 予算, その他の考慮事項に基づいて, 臨床試験全体の症例数を事前規定し, 検討を行う方法である。また, 事前規定した, 各 period の sub-sample の症例が集まった段階で, 安全性, 及び/または有効性の閾値を越えるかなどの判断を行う方法である。Probability monitoring procedure には, non-adaptive と adaptive の Probability monitoring procedure がある。Adaptive の Probability monitoring procedure では, 安全性, 無益性, 及び/または有効性の基準に基づいて, 試験を途中で中止することができる。そのため, 本節では Adaptive probability monitoring procedure についてのみ説明を行う。Probability monitoring procedure では, 以下の 1 標本の割合に関する仮説を考えていく。すなわち, 罹患率が低い方が好ましい状況を考える。

$$H_0 : p \geq p_0 \quad \text{versus} \quad H_a : p < p_0$$

p_0 は事前に設定された, 臨床的に意味のある罹患率の閾値である。次節では, Adaptive probability monitoring procedure の手順について説明を行う。

3.6.3.3.1 Adaptive probability monitoring procedure の手順と解析手法

以下の手順で実施する。

i. 実施可能性, 予算, その他の考慮事項に基づいて, 臨床試験全体の症例数, 及び各 period の sub-sample を事前に設定する。例えば, 全体の症例数を$N = 800$, period を$Q = 8$とする。このとき, 各 period の累積の sub-sample は$\{s_1, s_2, \cdots, s_Q\}$となり,

それぞれ$n_1, n_2, \cdots, n_q, s_1 \subset s_2 \subset \cdots \subset s_Q$とすると各 period で 100 例ずつ集積する状況を考え$n_1 = 100, n_2 = 200, \cdots, n_Q = n = 800$となる。

ii. 無益性中止のための閾値P_{f_q}, 及び/または有効性中止のための閾値であるP_{e_q}を設定する。ただし，$q = 1, \cdots, Q$とする。第 1 種の過誤確率の超過を制御するために, MIP, MSP, MPP などの考え方に基づいて閾値を決定することができる（Chow and Chang, 2011）。

iii. sub-sample のs_qに対する発現率をr_qとし，二項分布の累積分布関数に基づく確率$P_q = B(r_q; n_q, p_q)$を計算する。P_qは s_qから観察された罹患率に基づいて推定される。最も妥当な推定値の 1 つとして下記が考えられる。$\hat{p}_{q-1} = \frac{r_{q-1}}{n_{q-1}}$, $p_1 = p_0$, φは$\varphi = 0.025$のように事前に規定された値とする。

$$P_q = \min\left(p_0, \hat{p}_{q-1} + z_\varphi \sqrt{\frac{\hat{P}_{q-1}(1-\hat{p}_{q-1})}{n_{q-1}}}\right)$$

$P_q < P_{e_q}$の場合は有効とみなして試験を有効中止し，$P_q < P_{f_q}$の場合は無効中止を行う。その他の場合は，試験を継続し，試験が早期中止になるか，試験が完了するまで iii.の手順を繰り返す。

3.6.3.3.2 Adaptive probability monitoring procedure の数値例，及び考察

以下の設定で Adaptive probability monitoring procedure を行うことを考える。

希少疾患における，現在の既存の治療法における罹患率をp_0＝20%または 50%と想定する。新規の治療法における罹患率が 20%または 50%から下がる方向に改善することを期待して，薬剤の効果を検討する単群試験を計画する。このとき，現実的に集められる最大の症例数が N=60, 600 とし, period は Q=3, 各 sub-sample（$n_1, n_2, \cdots, n_Q, s_1 \subset s_2 \subset \cdots \subset s_Q$）で同じ閾値$P_{f_q} = 0.2$に基づいて無益性のために早期中止する状況を考える。各 sub-sample のs_qで観測された発現数をr_qとし，対応する累積確率を P_qとする。P_qはs_qから観察された罹患率に基づいて下記の式によって推定される。ただし，$\hat{p}_{q-1} = r_{q-1}/n_{q-1}$, $p_1 = p_0$, $\varphi = 0.025$とする。

$$P_q = \min\left(p_0, \hat{p}_{q-1} + z_\varphi \sqrt{\frac{\hat{P}_{q-1}(1-\hat{p}_{q-1})}{n_{q-1}}}\right)$$

		$p_0 = 0.2, P_f = 0.20$				$p_0 = 0.5, P_f = 0.20$			
Q	n_q	P_q	r_q	p_q	r_q/n_q	P_q	r_q	p_q	r_q/n_q
1	20	0.069	1	0.2	0.05	0.132	7	0.5	0.35

2	40	0.147	3	0.15	0.08	0.134	16	0.5	0.40
3	60	0.157	8	0.2	0.13	0.183	26	0.5	0.43

		$p_0 = 0.2, P_f = 0.20$				$p_0 = 0.5, P_f = 0.20$			
Q	n_q	P_q	r_q	p_q	r_q/n_q	P_q	r_q	p_q	r_q/n_q
1	200	0.166	34	0.2	0.17	0.179	93	0.5	0.47
2	400	0.175	72	0.2	0.18	0.198	191	0.5	0.48
3	600	0.194	111	0.2	0.19	0.196	289	0.5	0.48

上記は，それぞれの条件の下でP_qがP_fを上回らないときの最大のr_qを示している。例えば，$p_0 = 0.2,\ P_f = 0.20,\ N = 60,\ n_1 = 20$のとき発現数$r_1$が 1 だったとする。このとき，対応する累積確率は P_1=0.069 となり$P_f = 0.20$を上回らない。そのため，試験を継続する判断を下す。

r_q/n_qの結果から症例数が少ない場合には，症例数が多い場合に比べ，試験を中止する判断が下されやすくなる。例えば，$p_0 = 0.2,\ P_f = 0.20,\ n_1 = 20$のとき$r_q/n_q = 0.05,\ n_3 = 60$のときは$r_q/n_q = 0.13$となる。つまり，$n_1 = 20$のときは$n_3 = 60$のときに比べ，試験を中止する判断が下されやすくなる。また，r_q/n_qの結果からp_0が 0.5 から離れた値の場合は，0.5 付近の値に比べ試験を中止する判断が下されやすくなる。例えば，$p_0 = 0.2,\ P_f = 0.20,\ n_1 = 20$のとき$r_q/n_q = 0.05,\ p_0 = 0.5, P_f = 0.20, n_1 = 20$のときは$r_q/n_q = 0.35$となる。つまり，$p_0 = 0.2$のときは$p_0 = 0.5$のときに比べ，試験を中止する判断がされやすくなる。そのため，モニタリングの早期段階や治療法における罹患率が 0.5 から離れた値では，症例数が多い場合や治療法における罹患率が 0.5 付近の値に比べ，厳しめの無益性の閾値の設定は避ける必要があるかもしれない。

Probability monitoring procedure は，実施可能性，予算，その他の考慮事項に基づいて，事前に実施可能な症例数に基づいて検討を行う方法である。そのため，検出力や推定精度に基づいた臨床試験を実施できない場合にも適用可能である。また，Adaptive monitoring procedure は各 sub-sample ごとに期待される臨床効果を更新し，安全性，無益性，及び／または有効性に基づいて試験を早期に中止することができる。そのため，複雑であるが適切な閾値を設定できれば，時間とリソースを節約できる可能性がある。例えば，希少疾患などで症例登録可能な患者が少ない場合に，探索的に薬剤の効果を観察しながら開発を進めていくような状況であれば，ある程度緩い安全性/有効性/無益性の閾値を設定することで，このような方法を用いる余地があるかもしれない。検証的試験で用いるような有意水準の設定での適用は難しい可能性がある。また，本報告書で説明した Probability monitoring procedure は，アダプティブデザインなどの複雑なデザインにも適用可能である（Huang 2019[1]）。

参考文献

[1] Huang, Zhipeng, and Shein-Chung Chow. “Probability monitoring procedures for sample size determination.” Journal of Biopharmaceutical Statistics 29.5 (2019): 887-896.

3.7　有意水準を両側 5%超で設定した試験

3.7.1　概要

本節では，検証的試験において両側有意水準 5%（または片側有意水準 2.5%）よりも大きな有意水準を採用した事例を取り上げる。ICH-E9 ガイドライン「臨床試験のための統計的原則について」[1]には，「検証的位置づけの試験を行う際の有意水準（第一種の過誤）については従来明確にされていなかったが，規制上の観点から，本ガイドラインの施行に伴い，原則として片側仮説を検証する場合は 2.5%，両側仮説の場合は 5%とすることとした」との記載がある。

一方で，「臨床試験のための統計的原則」に関する質疑応答では，「適切な説明ができるのであれば，より強固な有効性の根拠を示すために有意水準を厳しくする，希少疾病用医薬品にみられる例のように十分な被験者を集めることが困難な場合は有意水準を緩くする，などの措置をとってもよい。」との記載がある。また，FDA から 2019 年に発出された Demonstrating Substantial Evidence of Effectiveness for Human Drug and Biological Products [2]においても，治療法が存在しない深刻な疾患や希少疾患のような症例数が限られる場合においては，事前に規定され適切な正当化がされることを条件に，0.05 よりも大きい p 値が許容される可能性があるとの記載がある。

患者数が極めて限られる状況においては両側有意水準 5%（または片側有意水準 2.5%）から算出される必要症例数を登録することに大きな困難が生じる可能性がある。このような場合においては，有意水準の設定に限定されるものではないが，試験開始前に治験相談を活用して，科学的な評価が可能で，かつ実施可能性のある試験デザインについて，規制当局と合意を得ることが望まれる。

なお，抗悪性腫瘍薬における希少がんや希少フラクションを適応症とした申請において，片側有意水準 5%（両側 90%信頼区間）を用いた単群試験の結果で承認申請されている例が散見されるが，このような場合も，試験開始前に有意水準について規制当局と合意を形成することが望ましいと考える。

3.7.2　リュープロレリン酢酸塩（リュープリン®SR 注射用キット 11.25mg）

<u>リュープロレリン酢酸塩について</u>

本薬剤の対象疾患である球脊髄性筋萎縮症は，成人男性のみに発症する下位運動ニューロン疾患であり，四肢近位部の筋力低下・筋萎縮と球麻痺を主症状とする。筋力低下の発症は通常 30～60 歳頃であり，緩徐進行性の経過を辿り，末期には患者は車椅子上または寝たきりの生活を余儀なくされるばかりではなく，誤嚥性肺炎を繰り返しこれが死因になることが多い。国内での有病率は 10 万人あたり 1～2 人程度と推計される。また，国内外において，球脊髄性筋萎縮症に対する有効な治療法は確立しておらず，治療法は存在しない。

リュープロレリン酢酸塩は，高活性の黄体形成ホルモン放出ホルモンアゴニストであり，性腺機能を抑制し，性ステロイドホルモンの分泌を低下させる薬剤である。国内においては，前立腺癌，閉経前乳癌等の性ホルモン依存性疾患の治療薬として承認されている。球脊髄性筋萎縮症の発症，及び進行の機序に，性ステロイドであるテストステロンが関与していることから，球脊髄性筋萎

縮症の効能追加について臨床開発され，2017 年に承認された。

リュープロレリン酢酸塩の臨床試験としては，日本人の球脊髄性筋萎縮症患者を対象とした国内第 III 相試験（06DB 試験）とその長期継続投与試験（07OP 試験），並びに追加国内第 II 相試験（11DB 試験）が評価資料であった（図 3-7-2-1）。この中で，国内第 III 相試験（06DB 試験）について，両側有意水準 10%が適用されている。

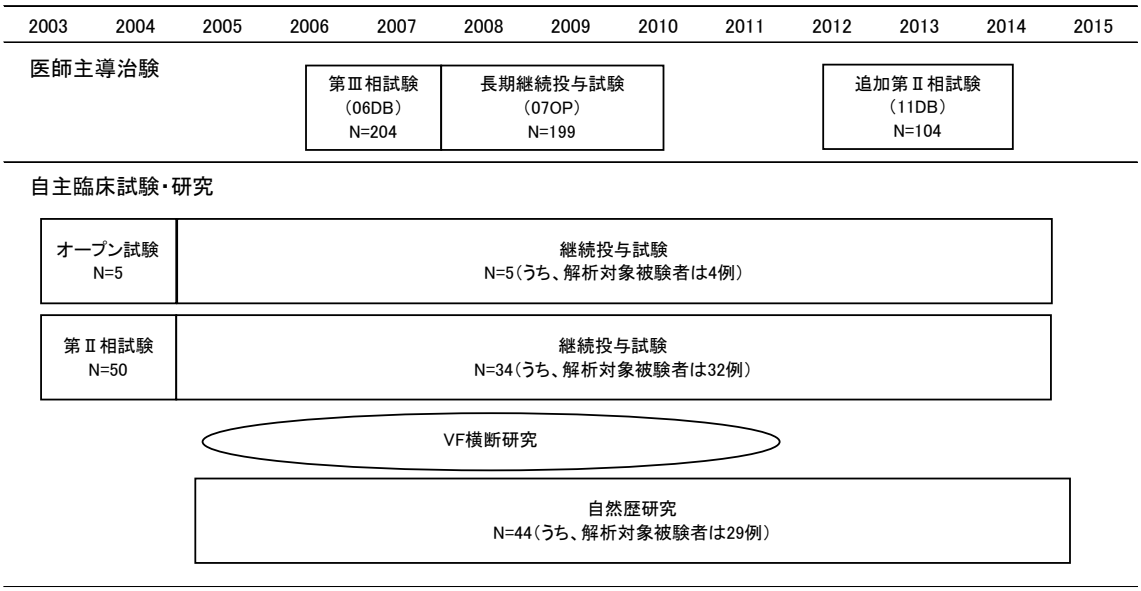


図 3-7-2-1　　臨床開発の概要

申請資料概要をもとに作成

<u>両側有意水準 10%が適用されている試験：06DB 試験</u>

06DB 試験は，日本人の球脊髄性筋萎縮症患者（目標症例数 170 例，各群 85 例）を対象に，リュープロレリン酢酸塩の有効性，及び安全性を確認するためのプラセボ対照ランダム化二重盲検並行群間比較試験である。被験者は，実薬群（11.25mg）またはプラセボ群に 1:1 で割付された。主要評価項目は，①旧 JASMITT（the Japan SBMA Interventional Trials for TAP-144-SR，なお SBMA は球脊髄性筋萎縮症を意味する）方式（50%以上除く），及び②新 JASMITT 方式による，投与終了時（投与開始 48 週後または中止時）の咽頭部バリウム残留率のベースラインからの変化量である。球脊髄性筋萎縮症患者では，嚥下障害の進行に伴い誤嚥性肺炎を繰り返し，誤嚥性肺炎が最終的な死因になる場合が多いとされていることより，嚥下機能に関連する指標として咽頭部バリウム残留率が主要評価項目として設定された。JASMITT 方式は，40%W/V の硫酸バリウム（以下ではバリウムと記載する）3mL を嚥下したとき，バリウムが食道に移行した時点での咽頭蓋谷，梨状陥凹へのバリウムの残留率を 10%刻みで目算する評価方法である。リュープロレリン酢酸塩の臨床試験では①から③の測定方法が用いられた。

①　旧 JASMITT 方式（50%以上含む）：初回嚥下時の残留率のみを評価する方式。すべての被

験者を解析対象集団に含める。

② 旧 JASMITT 方式(50%以上除く):初回嚥下時の残留率のみを評価する方式。球脊髄性筋萎縮症患者では嚥下機能が低下しているため，複数回に分けて小刻みに嚥下する特徴があり，被験者によっては初回嚥下が「主たる嚥下」とはならない場合があることから，ベースライン時及び治験薬投与後(24 週時，48 週時又は中止時のいずれか)で，初回嚥下時の残留率が50%未満である被験者(初回嚥下が「主たる嚥下」である被験者)のみを解析対象とする。

③ 新 JASMITT 方式:球脊髄性筋萎縮症患者に特徴的な複数回の小刻みな一連の嚥下全体を「主たる嚥下」としたときの残留率。

主要評価項目に関する試験結果を表 3-7-2-1 に示す。旧 JASMITT 方式(50%以上除く)を用いた評価では統計学的有意差は得られず(p=0.331)，一方で，新 JASMITT 方式を用いた評価では統計学的有意差が得られた(p=0.049)。

表 3-7-2-1 旧 JASMITT 方式(50%以上除く)，及び新 JASMITT 方式による投与終了時の咽頭部バリウム残留率(%)のベースラインからの変化量(FAS)

残留率の測定方法	投与群	評価例数 a)	測定値		ベースラインからの変化量	プラセボ群との比較	
			ベースライン	投与終了時		群間差[90%信頼区間]	p 値 b)
旧 JASMITT 方式(50%以上除く)	プラセボ群	77	7.77±8.28	8.03±7.18	0.25±7.16	-1.27 [-3.41, 0.88]	0.331
	実薬群	79	9.17±10.37	8.16±7.35	-1.01±8.93		
新 JASMITT 方式	プラセボ群	96	6.68±7.19	8.34±11.17	1.66±9.32	-3.21 [-5.89, -0.52]	0.049
	実薬群	97	10.58±13.53	9.04±11.30	-1.55±12.94		

平均値±標準偏差

a) FAS のうちベースライン時または試験薬投与後に測定結果が得られなかった被験者は解析から除外した。旧 JASMITT 方式(50%以上除く)の評価では，さらにベースラインまたは試験薬投与後の評価において初回嚥下時の残留率が 50%以上である被験者解析から除外した。

b) 2 標本 t 検定，有意水準は両側 10%(2 つの評価項目に対する検定の多重性は調整されていない)。

審査報告書 [3] の表 1 をもとに作成

ただし，主要評価項目の 1 つである，新 JASMITT 方式による投与終了時（投与開始 48 週後または中止時）の咽頭部バリウム残留率に関して，残留率のベースライン値に群間差が認められたことより，ベースライン値で調整した解析を実施した。その結果（表 3-7-2-2），実薬群とプラセボ群の間に統計学的な有意差は認められなかった（p=0.392，ベースライン値を共変量とした共分散分析モデルに基づく）。

表 3-7-2-2 新 JASMITT 方式による投与終了時の咽頭部バリウム残留率（%）のベースラインからの変化量（FAS）

投与群	評価例数 a)	測定値		ベースラインからの変化量 b) c)	プラセボ群との比較 c)	
		ベースライン	投与終了時		群間差［90%信頼区間］	p 値
プラセボ群	96	6.68±7.19	8.34±11.17	0.67±1.02	-1.24 [-3.63, 1.15]	0.392
実薬群	97	10.58±13.53	9.04±11.30	-0.57±1.01		

平均値±標準偏差

a) FAS のうちベースライン時または試験薬投与後に測定結果が得られなかった被験者は解析から除外した。

b) 調整平均値±標準誤差

c) 投与群を因子，ベースライン値を共変量とした共分散分析モデルに基づく。

審査報告書 [3] の表 2 をもとに作成

一方で，試験開始時に主要評価項目に設定されていた，旧 JASMITT 方式（50%以上含む）による投与終了時（投与開始 48 週後または中止時）の咽頭部バリウム残留率のベースラインからの変化量の結果は表 3-7-2-3 の通りであり，探索的な解析であるものの，実薬群とプラセボ群の間に，有意水準両側 10%で，統計学的な有意差が認められた（p=0.063，2 標本 t 検定）。06DB 試験では，球脊髄性筋萎縮症の病態や盲検化レビューの結果等も考慮して，試験開始後に主要評価項目である咽頭部バリウム残留率の測定方法を旧 JASMITT 方式（50%以上含む）から旧 JASMITT 方式（50%以上除く），及び新 JASMITT 方式に変更した経緯がある。

表 3-7-2-3 旧 JASMITT 方式（50%以上含む）による投与終了時の咽頭部バリウム残留率（%）のベースラインからの変化量（FAS）

投与群	評価例数 a)	測定値		ベースラインからの変化量	プラセボ群との比較	
		ベースライン	投与終了時		群間差［90%信頼区間］	p 値 b)
プラセボ群	96	18.65±26.64	18.83±24.56	0.18±18.15	-5.26 [-9.92, -0.60]	0.063
実薬群	98	20.30±27.08	15.22±20.37	-5.08±20.96		

平均値±標準偏差

a) FAS のうちベースライン時または試験薬投与後に測定結果が得られなかった被験者は解析から除外した。

b) 2 標本 t 検定，有意水準は両側 10%

審査報告書 [3] の表 3 をもとに作成

PMDA での審査について

リュープロレリン酢酸塩については，咽頭部バリウム残留率の変化量を主要評価項目として選択したことの適切性や，咽頭部バリウム残留率の評価方法の適切性，嚥下機能に対するリュープロレリン酢酸塩の有効性などについて，審査報告書で詳細に議論されている。結果として，06DB 試験の成績のみに基づいてリュープロレリン酢酸塩の有効性が示されたと判断することはできないと考えられたため，咽頭部バリウム残留率の測定・解析方法の適切性，及び 06DB 試験の再現性確認を目的として，追加試験（11DB）が実施された。

11DB 試験については，06DB 試験で得られた結果に基づくと，有意水準を両側 10%，検出力を 80%とした場合であっても症例数は各群 173 例となり，患者を集積することは困難と考えられた。そのため，実施可能性に基づいて目標症例数が設定されており，各群 40 例の探索的試験として実施された。11DB 試験では，統計学的な有意差は認められなかったものの，旧 JASMITT 方式（50%以上含む）による咽頭部バリウム残留率の変化量についてリュープロレリン酢酸塩の改善傾向が確認された。

最終的には，承認条件として，「国内での治験症例が極めて限られていることから，製造販売後，一定の症例に係るデータを集積されるまでの間は，全症例を対象とした使用成績調査を実施することにより，本剤使用患者の背景情報を把握するとともに，本剤の安全性，及び有効性に関するデータを早期に収集し，本剤の適正使用に必要な措置を講じること。」という条件が付与された上で，リュープロレリン酢酸塩の効能が追加された。

本事例では，06DB 試験において両側 10%という有意水準が適用されており，審査報告書には以下の記載がある。

本剤の検証的試験（06DB 試験）において有意水準が両側 10%と設定された[5]経緯は，明確化される必要があるとの意見が専門委員から示された。

この点について申請者は，06DB 試験の計画時に治験実施者が検討したところ，有意水準を両側 5%，検出力を 90%と設定した場合，中止率を考慮しない場合でも 1 群 128 例（計 256 例）の登録が必要になり，国内患者数（平成 10 年度実施の厚生省特定疾患調査研究事業により 830 人）を考慮すると試験の実施は困難と考えられたことから，やむを得ず有意水準を両側 10%と設定したと説明している。

機構は，臨床試験の実施可能性を考慮すると，有効性について一定の説明を行う上で有意水準を両側 10%としたことはやむを得なかったと考える。

まとめ

検証的試験で両側 5%または片側 2.5%よりも大きい有意水準が適用された事例として，リュープロレリン酢酸塩の事例を取り上げた。希少疾患を対象としており症例集積が非常に困難である臨床試験については，有意水準の設定に限定するものではないが，治験相談を活用して試験開始前に規制当局と相談することが望ましいと考える。本事例については，全体を通して，リュープロレリン酢酸塩の申請資料概要，及び審査報告書を参考にした。

参考文献

[1] 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.” 厚生省（1998）.

[2] Food and Drug Administration. "Demonstrating Substantial Evidence of Effectiveness for Human Drug and Biological Products -Guidance for Industry." (2019).

[3] 医薬品医療機器総合機構. "リュープリン SR 注射用キット 11.25mg, 審査報告書."（2017）. http://www.pmda.go.jp/drugs/2017/P20170824001/400256000_22700AMX00128_A100_1.pdf

5 第 II 相試験として実施された臨床試験（参考 CTD 5.3.5.1-4: Ann Neurol 2009; 65: 140-50）の成績を参考にプラセボに対するリュープロレリン酢酸塩の優越性を検証することが計画されたが，球脊髄性筋萎縮症は希少疾患であり，症例集積が非常に困難と考えられたため，有意水準は両側 10%と設定された。なお 2 つの主要評価項目に対する検定の多重性の調整は計画されていなかった。

3.8 代替評価項目の活用

3.8.1 概要

代替評価項目（代替エンドポイント, surrogate endpoint; SE）は臨床試験で評価されるアウトカムの測定の代わりになるものである。それ自体は臨床上のベネフィットを直接測定するものでないが, 臨床上のベネフィットを予測することが知られているバイオマーカー（臨床検査値, X 線診断画像, 身体的特徴, またはその他の測定など）が SE になり得る。SE が臨床アウトカムの代わりとして容認されることを示す広範なエビデンスがなければならない。あまり確立されていない SE であっても臨床上の有益性を予測する可能性が合理的に高いと考えられるものは, 重篤な疾患や生命を脅かす疾患の治療のための迅速承認の根拠として使用することができる可能性がある[1]。

規制上の観点から, SE は臨床的バリデーションのレベルにより以下のように分類できる[2]。

バリデートされた代替エンドポイント（Validated SE）

臨床アウトカムを高い信頼性で予測することができる SE で, 通常承認の根拠として用いることができる[3]。SE が改善されることはすなわち特異的な臨床上のベネフィットであることが, 作用機序に関する明確な論拠と臨床データによる強力なエビデンスによって裏付けられている。

妥当と考えられる代替エンドポイント（Reasonably likely SE）

その名が示す通り, 臨床上のベネフィットを予測できると考えられる SE である。これらの SE は, 有力な作用機序, 及び/または疫学にもとづく強力な論拠による裏付けはあるが, バリデートするために利用できる臨床データの量が不十分である。迅速承認（accelerated approval）の根拠として用いることはできるが, これらの SE が臨床上のベネフィットを予測するかまたは臨床上のベネフィットに相関することを高い信頼度で示すためには, 承認後の臨床試験が必要である。

代替エンドポイント候補（Candidate SE）

臨床上のベネフィットを予測できるかどうかについてまだ評価中の SE である。

新薬, 及び既存薬の新しい適応について臨床試験を行う際に, これらの SE を用いることができる。適切な試験において SE で有益な結果が示されれば, SE を使用して少数の患者で短期間の臨床試験の実施が認められ, それにより医薬品開発が迅速化する可能性がある。臨床アウトカムが有効性を直接測定するのに対し, SE は臨床アウトカムを予測する代替物を測定する。SE は, 特定の疾患の治療効果を見届けるのに非常に長期間を要する場合, あるいは SE の改善による臨床上のベネフィットが十分に理解されている場合には, 優れた代替手段となり得る。また, エンドポイントに関する臨床試験を行うことが非倫理的である場合にも, SE が使用されることがある。

FDA は医薬品開発での SE の使用を強化し, 患者のための革新的新薬の開発を促進するため, SE の使用を考えている医薬品開発プログラムが利用できる新たなリソースの提供を開始した[2]。

そのようなリソースの一つが，通常承認と迅速承認の両方についての「医薬品承認の根拠となった代替評価項目の一覧表」[4]である。

本項ではFDAの「医薬品承認の根拠となった代替評価項目の一覧表」(表3-8-1-1)にも記載のあるSEを使用して承認に至ったアガルシダーゼ ベータについて言及する。

表3-8-1-1 医薬品承認の根拠となった代替評価項目の一覧表[4]より一部抜粋

適応症/使用目的	患者集団	代替エンドポイント	承認のタイプ	作用機序
ファブリー病	ファブリー病と診断された患者	腎血管内皮細胞からの蓄積GL-3の完全/ほぼ正常レベルまでの除去(Fabrazyme スコアリングシステムを使用)	通常型	酵素補充療法

3.8.2 アガルシダーゼ ベータ(ファブラザイム®)

ファブリー病について

ファブリー病はリソソームの加水分解酵素であるα-ガラクトシダーゼ A(以下αGAL)の活性が正常以下または欠損していることを特徴とするX染色体劣性の先天性代謝異常である。αGALが欠乏すると，血管の内皮細胞，外膜，及び中膜平滑筋のリソソーム中にスフィンゴ糖脂質，主にグロボトリアオシルセラミド(GL-3)が進行性に蓄積する。GL-3の蓄積は，自律神経節細胞，心筋細胞，糸球体上皮細胞，尿細管，及び角膜等で生じる。典型的(古典的)ファブリー病男性患者(ホモ接合体)ではαGAL活性がほぼ完全に欠損し，幼少時から四肢の激痛(先端異常感覚)の散発的な発症，特徴的な皮膚病変の出現，発汗低下，特徴的な角膜，及び水晶体の混濁などの症状が出現し，やがて臓器の機能不全へと進行する。加齢とともに腎不全，心臓疾患または脳血管障害を生じて死亡する。

日本では1994年に実施された調査[5]においてファブリー病患者138名が報告された。当時，ファブリー病の原因の軽減ないし進行の抑制を期待できる確立した方法はなかったため，対症療法が行われており，疼痛や持続的な不快感を軽減するための疼痛管理が主な薬物治療であった。また，透析と腎移植が末期腎不全に対する治療の選択肢であった。

アガルシダーゼ ベータについて

アガルシダーゼ ベータは，ファブリー病の酵素補充療法を目的として開発され，ヒト線維芽細胞のmRNAに由来するヒトαGALの相補的DNAを，遺伝子組換え技術によりチャイニーズハムスター卵巣細胞(CHO細胞)に組み込んで産生される，αGALである。2001年8月に欧州，2003年4月に米国，その他13カ国で承認されており，日本においては2004年に承認された。

本事例は客観的な評価項目である腎生検組織におけるGL-3の除去を代替評価項目として選定

し承認された事例として紹介する。

代替評価項目の設定

腎不全はファブリー病の進行，及び死亡の主な原因の 1 つである。しかし，腎機能が正常な患者集団について腎機能の維持を評価項目とする試験を実施するには，多数の患者を対象とした長期間の試験が必要である。患者数が極めて少ない本疾患において臨床的有効性を評価項目とした臨床試験を実施することは困難と考えられたため，腎血管内皮細胞からの蓄積 GL-3 のほぼ正常レベルまでの除去という代替評価項目を主要評価項目とした。

試験結果

有効性，及び安全性の評価については，国内第II相試験（AGAL-007-99）と海外における第I/II相試験（FB9702-01）で，ファブリー病患者における隔週 11 回（20 週間）の投与により血漿，尿，腎，心臓，及び皮膚組織中に蓄積している GL-3 の除去が認められた。国内第II相試験（AGAL-007-99）においては，日本人と欧米人の間で有効性と安全性について比較することも目的とされており，有効性と安全性の両方について全ての臨床的なエンドポイントを評価するように計画され，主要評価項目，副次評価項目等に分類はされなかった。腎臓病理では，腎臓組織標本中の毛細血管内皮細胞の GL-3 蓄積スコア（0:なし，1:軽度， 2:中等度，3:重度）で評価され，13 例中 10 例がベースラインのスコア 1 から 11 回目投与後（第 20 週目）に 0 へ，残る 3 例のベースラインはスコア 2 であったが，内 2 例が 0 へ，1 例が 1 へ減少し，ベースラインと 20 週目でのスコア 0 の比率には有意な差が認められた（$p<0.001$，Wilcoxon 符号付順位和検定）。また，腎臓，皮膚，心臓において GL-3 の除去が認められた(表 3-8-2-2)。

表 3-8-2-2 組織学的評価による毛細血管内皮細胞の蓄積 GL-3 の除去効果（試験終了時におけるスコアゼロ※の達成数）

	日本における第 II 相試験 AGAL-007-99	第 III 相二重盲検比較試験（5 か月間） AGAL-1-002-98		第 III 相非盲検継続試験 AGAL-005-99			
				（6 か月間）		（54 か月間）	
		プラセボ群	実薬群	プラセボ／実薬群	実薬／実薬群	プラセボ／実薬群	実薬／実薬群
腎臓	12/13	0/29	20/29	24/24	23/25	5/5	3/3
心臓	1/1	1/29	21/29	13/18	19/22	3/5	3/3

	日本における第II相試験 AGAL-007-99	第III相二重盲検比較試験(5か月間) AGAL-1-002-98		第III相非盲検継続試験 AGAL-005-99			
				(6か月間)		(54か月間)	
		プラセボ群	実薬群	プラセボ／実薬群	実薬／実薬群	プラセボ／実薬群	実薬／実薬群
皮膚	12/13	1/29	29/29	25/26	26/27	17/19	14/17

※スコアゼロ:細胞中に蓄積物質の封入体が認められないか痕跡程度[5]

アガルシダーゼ ベータの安全性及び有効性を評価する目的で施行された海外における第III相二重盲検比較試験(AGAL-1-002-98)では, 有効性について腎組織病理では, 5か月目時点でアガルシダーゼ ベータ投与の継続によりGL-3が消失し維持されることが示された[6]。

また, アガルシダーゼ ベータの長期有効性及び安全性パラメータについて評価する目的で, 多国間, 多施設, オープン(非盲検非対照)試験(AGAL-1-002-98)が実施された。本試験の対象者はAGAL-1-002-98に登録された症例全58例がこの継続試験に参加した。最長54か月間の投薬で腎臓及び皮膚の様々な細胞でGL-3除去効果が認められた[6]。

審査上の論点

本試験では, ファブリー病患者の組織中に蓄積した GL-3 の除去効果について検討されているが, ファブリー病の臨床症状に対する改善効果については明確に示されていない。しかし, アガルシダーゼ ベータが患者数の少ない希少疾病を対象にしているため, 腎不全の進行抑制等, 真のエンドポイントを主要評価項目として試験を実施する事が困難であったことを了承し, 代替評価項目に組織(腎臓の毛細血管内皮)中の GL-3 除去を選定したこともFDA との協議の経緯に鑑み是認された。この主要評価項目で合意に至った理由は以下の通りである。

① 腎不全はファブリー病で最も共通した症状である。
② ファブリー病における糸球体硬化症の主原因は腎血管内皮細胞への GL-3 蓄積による血管障害にある。
③ この評価項目は合理的な期間内に評価可能である。
④ この評価項目に対する統計学的検出力の算定より, 患者数の極めて少ない本疾患でも適切な症例数の試験が実施可能である。
⑤ FDA, 及び当該分野の専門家との協議により, 腎毛細血管内皮細胞からの GL-3 のほぼ正常レベルまでの除去は臨床上重要であり, 機能の正常化, 及び臨床効果を予測し得ると考えられた。
⑥ ファブリー病による腎不全では, 治療が必要となるまで明確な臨床症状が現れないため, 組織学的診断が疾患の程度, 及び治療の客観的指標であることが少なくない。

上記より機構は，患者数の少ない本疾患で比較的短期間の試験により判定し得る客観的な評価項目として腎生検組織における GL-3 の除去を代替評価項目として選定した理由と経緯の説明を了承した。アガルシダーゼ ベータの投与により組織中の GL-3 抑制効果は認められたものの，臨床症状の改善効果については確認されていないことから，多施設共同ランダム化プラセボ対照二重盲検の第IV相試験（AGAL-008-00）が海外で行われ[7]，日本においても全投与症例を対象とした市販後調査，長期投与による特別調査により情報を収集することが承認条件とされた[5]。

参考文献

[1] Food and Drug Administration. "Rare Diseases: Common Issues in Drug Development - Guidance for Industry (Draft)." (2019).

[2] 国立医薬品食品衛生研究所．医薬品安全情報 2019;17:2-13. http://www.nihs.go.jp/dig/sireport/weekly17/01190110.pdf

[3] FDA Facilitates the Use of Surrogate Endpoints in Drug Development" November 5, 2018 Issue. https://www.fda.gov/drugs/fda-facilitates-use-surrogate-endpoints-drug-development-november-5-2018-issue

[4] Food and Drug Administration. “Table of Surrogate Endpoints That Were the Basis of Drug Approval or Licensure.” https://www.fda.gov/drugs/development-resources/table-surrogate-endpoints-were-basis-drug-approval-or-licensure.

[5] 医薬品医療機器総合機構．"ファブラザイム，審査報告書." （2003）．https://www.pmda.go.jp/drugs/2004/P200400006/34053100_21600AMY00008_A100_1.pdf

[6] ファブラザイム®添付文書

[7] Banikazemi, Maryam, et al. "Agalsidase-beta therapy for advanced Fabry disease: a randomized trial." Annals of internal medicine 146.2 (2007): 77-86.

4 まとめ

希少疾病用医薬品の臨床試験では，これまでにも述べた通り患者数が少ない又は倫理的な面からプラセボ同時対照が設定できないなどの理由から標準的なランダム化比較試験の下で治療効果を精度高く推測することが困難な場合がある。しかしながら，このような状況下でも一定の水準で有効性及び安全性のエビデンスを創出する必要があることから，自然歴研究などを開発初期に実施することで疾患に関する情報を収集しその後の開発に活用すること，及び革新的な試験デザインと対応する統計手法が考慮されるべきと考えられる。革新的な試験デザインとは症例数低減や試験期間短縮を実現すると同時に治療効果推測の精度を保つ又は精度を高めるものである。この点は希少疾患を患っている患者様のアンメットメディカルニーズに応える面からも重要である。

本報告書では，試験途中のアダプテーションを伴う試験における第 1 種の過誤確率の制御や，ヒストリカルコントロール等の有効性に関する事前情報と試験で観測されたデータを統計的に統合するベイズ流アプローチなどを取り上げ，統計手法の解説とともに適用事例を示した。適用事例には可能な限り審査報告書を参照し規制当局の見解を含めた。また，最近の議論として N-of-1 試験デザイン，Complete N-of-1 試験デザイン，Probability of inconclusive，及び Probability monitoring procedure も取り上げた。関連して，N-of-1 試験及び Complete N-of-1 試験に対する症例数設計法を補遺にまとめたので合わせて参照されたい。

本報告書をまとめる過程において，公表されている審査事例のため限定的であるが，効率的にデータを活用するための革新的な試験デザインや革新的な試験デザインに対応する統計手法を利用した医薬品開発が行われていたことが明らかになった。今後，革新的な試験デザイン及び統計手法に対する議論や情報共有が進めば，希少疾病用医薬品の臨床評価が将来的に更に進展する可能性があると言えるであろう。

革新的な試験デザイン及び統計手法は，症例数及び試験期間の面で効率化に寄与することが期待される半面，計画時にはバイアスや一般化可能性，そして選択した統計手法の動作特性などを十分に考慮する必要がある。また，希少疾病用医薬品の承認申請資料に革新的な試験デザイン及び統計手法（特にベイズ流アプローチ）に基づくエビデンスが含まれる場合，スポンサーと規制当局の間で事前にどのような合意事項が必要であるかの議論も今後深めていく必要があると考えられる。

本報告書が多くの方々の手に取られ，希少疾病用医薬品開発促進の一助となれば幸いである。

補遺

1. 群逐次法におけるファミリーワイズの第 1 種の過誤確率の制御

1.1 基本的な群逐次法及び過誤消費関数による方法

2 群比較試験において，各治療群からの測定値を$Y_{ij}\sim N\left(\mu_j,\sigma^2\right),i=1,\cdots,n_j,j=1,2$ とし，K回の解析を行うとする（K回目は最終解析。また，解析間の情報量の増加は等しい，すなわち，この例では試験の総症例数 Nに対してk回目の中間解析で用いられる症例数は$N_k=\frac{k\times N}{K}$ とする。）。$\theta=\mu_1-\mu_2$ について，$\mathrm{H}_0:\theta=0,\mathrm{H}_1:\theta\neq 0$ という帰無仮説，対立仮説を考える。この時，第k回目の中間解析に関する検定統計量は，それまでに蓄積されたデータによって$Z_k=\frac{1}{\sigma\sqrt{N_k}}\sum_{i=1}^{\frac{N_k}{2}}(Y_{i1}-Y_{i2}),k=1,\cdots,K$で表される。Pocock (1977)は，ファミリーワイズの第 1 種の過誤確率をα 以下に制御するため，$\Pr(\cup_{k=1}^{K}|Z_k|\geq C_{Po}(K,\alpha)\big|\mathrm{H}_0)=\alpha$ となる棄却係数$C_{Po}(K,\alpha)$を計算した[1]。これにより，各解析時に得られる$|Z_k|$が$C_{Po}(K,\alpha)$よりも大きければ帰無仮説を棄却し，そうでなければ試験を継続することになる。これに対し，O'Brien and Fleming (1979)は，$\Pr\left(\cup_{k=1}^{K}|Z_k|\geq C_{OF}(K,\alpha)\times\sqrt{\frac{K}{k}}\middle|\mathrm{H}_0\right)=\alpha$ を満たす棄却係数$C_{OF}(K,\alpha)$を算出した[2]。これにより，各解析時に得られる$|Z_k|$が$C_{OF}(K,\alpha)\times\sqrt{\frac{K}{k}}$ よりも大きければ帰無仮説を棄却し，そうでなければ試験を継続することになる。上述の Pocock の方法と異なり，早期の中間解析時には帰無仮説の棄却は非常に難しく，逆に試験が進むにつれて中止境界は緩和される。Wang and Tsiatis (1987) は，これらの手法も含む一般化した逐次検定の族を提案している[3]。すなわち，べき乗パラメータγを用いて，$\Pr\left(\cup_{k=1}^{K}|Z_k|\geq C_{WT}(K,\alpha,\gamma)\times\left(\frac{k}{K}\right)^{\gamma-\frac{1}{2}}\middle|\mathrm{H}_0\right)=\alpha$となる棄却係数$C_{WT}(K,\alpha,\gamma)$を算出した。$\gamma=0.5,0$の場合は，それぞれ Pocock，O'Brien and Fleming の手法と合致する。

更に，Lan and Demets (1983) は，第 1 種の過誤確率を情報時間の連続的な関数として消費するα消費関数を提案した[4]。これは，2 つの情報時間：$s_1,s_2,0<s_1<s_2\leq 1$としたときに，$0<\alpha(s_1)<\alpha(s_2)\leq\alpha$ を満たす関数であり，この$\alpha(s)$は時間s までに消費された第 1 種の過誤確率である。検定統計量$Z_k,k=1,\cdots,K$ に対して，帰無仮説の下で，第$(k-1)$回目まで帰無仮説を棄却せずに第k回目で帰無仮説を棄却する確率を，次のようにα消費関数を用いて表し，これを満たすように中止境界C_kを定める。

$$Pr(Z_1<C_1,Z_2<C_2,\cdots,Z_{k-1}<C_{k-1},Z_k\geq C_k)=\alpha(s_k)-\alpha(s_{k-1})=\alpha\left(\frac{k}{K}\right)-\alpha\left(\frac{k-1}{K}\right)$$

主なα消費関数としては以下のようなものがあり，上述の Pocock や O'Brien and Fleming の方法の棄却係数は，α消費関数により近似することができる[5]。

O'Brien and Fleming	$\alpha_1(s) = 2\{1 - \Phi(z_{\alpha/2}/\sqrt{s})\}$
Pocock	$\alpha_2(s) = \alpha \times log\{1 + (e-1)s\}$
Lan and Demets	$\alpha_3(s) = \alpha \times s^{\theta}, \theta > 0$
Hwang, Shih and DeCani	$\alpha_4(s) = \alpha \times \{(1 - e^{\zeta s})/(1 - e^{-\zeta})\}, \zeta \neq 0$

1.2 多段階デザイン（Bauer-Kohne 法）

解析時点までに蓄積されたデータに基づく p 値ではなく，解析時点間の Stage ごとのデータから算出する独立な p 値に基づく多段階のアダプティブデザインも検討されている。例えば，Bauer and Kohne (1994) は，Fisher の結合基準を用いた 2 段階のアダプティブデザインを提唱している[6]。試験を症例が重複しない 2 つの Stage に分け，それぞれのデータを同一の仮説を検定するために用いる。第 1，第 2 Stage における帰無仮説H_{01}，H_{02} とし，試験全体の帰無仮説を$H_0: H_{01} \cap H_{02}$で表す。各 Stage のデータから得られた p 値をp_1，p_2とすると，帰無仮説の下ではp_1，p_2は独立に一様分布に従う。Bauer-Kohne 法では，この性質を利用し，Fisher の結合基準に基づいて試験終了時に$p_1 p_2 \leq c_\alpha \equiv \exp(-\chi^2_{4,\alpha}/2)$ であれば帰無仮説H_0 は棄却される。ここで，$\chi^2_{4,\alpha}$は，自由度 4 のカイ二乗分布の$(1-\alpha)$パーセント点である。

具体的な判断規則としては，α_1，β_1 $(\alpha_1 < \beta_1)$ をカットオフ値としたとき，

第 1 Stage で$p_1 \leq \alpha_1$であれば，試験を中止し，H_0 は棄却する。

第 1 Stage で$p_1 > \beta_1$であれば，試験を中止し，H_0 は採択する。

第 1 Stage で$\alpha_1 < p_1 \leq \beta_1$であれば，試験を継続し第 2 Stage に進み，試験終了時に$p_1 p_2 \leq c_\alpha \equiv \exp(-\chi^2_{4,\alpha}/2)$ を満たせば，H_0 を棄却する。

なお，α_1，β_1 は，以下の式により算出する。p_1, p_2は帰無仮説の下で独立に一様分布に従うこと，Fisher の結合基準を用いることから，試験全体の第 1 種の過誤確率を左辺のように表現でき，この確率がαとなるように定式化している。

$$\alpha_1 + \int_{\alpha_1}^{\beta_1} \int_0^{\frac{c_\alpha}{p_1}} dp_2\, dp_1 = \alpha_1 + c_\alpha(log\beta_1 - log\alpha_1) = \alpha$$

ここで，Fisher の結合基準より，$c_\alpha = \exp(-\chi^2_{4,\alpha}/2)$の関係式を用いて，次式を得る。

$$\alpha_1 + (log\beta_1 - log\alpha_1)\exp(-\chi^2_{4,\alpha}/2) = \alpha$$

なお，Bouer-Kohne 法は，3 段階アダプティブデザインへの拡張が可能である[7]。

1.3 独立な p 値に基づく方法の一般化

独立な p 値に基づく方法については，より多段階のアダプティブデザインにも拡張できる[7, 8]。η_{k1}とη_{k2}をそれぞれk番目の Stage における各群のレスポンスとし，各 Stage の仮説を$H_{0k}: \eta_{k1} \geq \eta_{k2}, H_{ak}: \eta_{k1} < \eta_{k2}, k = 1, \cdots, K$ としたとき，試験全体の帰無仮説は以下のように表される。

$$H_0: H_{01} \cap, \cdots, \cap H_{0K}$$

各 Stage から得られる部分標本から得られる検定統計量，及び p 値を$T_k, p_k, k = 1, \cdots, K$ とし，以下の中止基準を考える。ここで，α_k, β_kはそれぞれ有効及び無効境界であり，$\alpha_k < \beta_k$ $(k = 1, \ldots, K-1), \alpha_K = \beta_K$である。

- 第k Stage で$T_k \leq \alpha_k$であれば，試験を中止し，H_0 は棄却する。
- 第k Stage で$T_k > \beta_k$であれば，試験を中止し，H_0 は採択する。
- 第k Stage で$\alpha_k < p_k \leq \beta_k$であれば，試験を継続する。

第k Stage の検定統計量T_kの累積分布関数は，第 1 Stage から第$(k-1)$ Stage で中止せずに試験を継続しなければならないことに注意すると，次式で与えられる。

$$\varphi_k(t) = P(T_k < t, \alpha_1 < t_1 < \beta_1, \ldots, \alpha_{k-1} < t_{k-1} < \beta_{k-1})$$
$$= \int_{\alpha_1}^{\beta_1} \ldots \int_{\alpha_{k-1}}^{\beta_{k-1}} \int_0^t f_{T_1 \ldots T_k} \, dt_k dt_{k-1} \ldots dt_1$$

$f_{T_1 \ldots T_k}$は，$T_1, \ldots, T_k$の同時確率密度関数である。第 k Stage での過誤確率（αの消費）は P($T_k < \alpha_k$)であるため，累積分布関数を用いて次のように表せる。

$$\pi_k = \varphi_k(\alpha_k)$$

試験全体の第 1 種の過誤確率は，各ステージの過誤確率の和として，次のようになる。

$$\alpha = \sum_{k=1}^{K} \pi_k$$

治療効果の強さを示す統計的指標として，調整済み p 値を算出できる。調整済み p 値は，第 k Stage で$T_k = t$ が得られた場合，以下のように定義される。

$$\mathrm{p}(t; k) = \sum_{i=1}^{k-1} \pi_i + \varphi_k(t), k = 1, \ldots, K$$

調整済み p 値は，試験全体の第 1 種の過誤確率αに関連し，式のとおり，より後半のステージでH_0が棄却されるほど，調整済み p 値は大きくなる。

続いて，検定統計量T_k の選択に関して解説する。これまでに様々なT_kが提案されており，例えば，個々の p 値に基づく方法（MIP: method of individual p-values）の他，p 値の和に基づく方法（MSP: method of sum of p-values），p 値の積に基づく方法（MPP: method of product of p-values）などがある。各方法のT_kは以下の式で与えられる。

$$T_k = p_k \qquad \text{(MIP)}$$
$$T_k = \sum_{i=1}^{k} w_{ki} p_i \, , w_{ki} > 0 \qquad \text{(MSP)}$$
$$T_k = \prod_{i=1}^{k} p_i \qquad \text{(MPP)}$$

調整 p 値の算出について，簡便のため$K = 2$とした場合について説明する。この時の試験全体の帰無仮説は:

$$H_0: H_{01} \cap H_{02}$$

また, この 1, 2 回目の解析で求められた粗 p 値をp_1, p_2 とする。この時, 帰無仮説の下では, p_1, p_2 はそれぞれ[0,1]の間で一様分布に従う。各 stage から得られた粗 p 値の統合には様々な提案法があるが, ここでは MSP（Method of Sum of p-value）として線形和 $T_k = \sum_{i=1}^{2} p_i$ を統計量とした方法を紹介する。

例えば, 第 Stage1 において, 早期有効中止のみを検討する場合は,

$$\pi_1 = \psi_1(\alpha_1) = \int_0^{\alpha_1} dt_1 = \alpha_1$$

$$\pi_2 = \psi_2(\alpha_2) = \int_{\alpha_1}^{\alpha_2} \int_{t_1}^{\alpha_2} dt_1\, dt_2 = \frac{1}{2}(\alpha_2 - \alpha_1)^2$$

これにより, $\alpha = \pi_1 + \pi_2 = \alpha_1 + \frac{1}{2}(\alpha_2 - \alpha_1)^2$であるから, $\alpha_2 = \sqrt{2(\alpha - \alpha_1)} + \alpha_1$となる。調整済み p 値は

$$\mathrm{p}(t; k) = \psi_1(t) = t = p_1$$

$$\mathrm{p}(t; 2) = \pi_1 + \psi_2(t) = \alpha_1 + \frac{1}{2}(t - \alpha_1)^2 = \alpha_1 + \frac{1}{2}(p_1 + p_2 - \alpha_1)^2$$

で表される。

なお, 早期有効中止に加えて早期無効中止も検討する場合, または早期無効中止のみ検討する場合でも, 同様に調整 p 値は導出が可能である[5,7]。早期有効中止, 及び早期無効中止を考慮する 2 段階アダプティブデザインにおいて, 検定統計量を MIP, MSP, MPP に基づき構築した場合の調整 p 値を以下に示す。

MIP	$p_1^* = p_1$ $p_2^* = \alpha_1 + (\beta_1 - \alpha_1)p_2$
MSP	$p_1^* = p_1$ $p_2^* = \alpha_1 + (p_1 + p_2)(\beta_1 - \alpha_1) - \frac{1}{2}(\beta_1^2 - \alpha_1^2), \beta_1 < \alpha_2,$ $p_2^* = \alpha_1 + \frac{1}{2}(p_1 + p_2 - \alpha_1)^2, \beta_1 \geq \alpha_2.$
MPP	$p_1^* = p_1$ $p_2^* = \alpha_1 + (p_1 p_2)\ln\left(\frac{\beta_1}{\alpha_1}\right), \beta_1 < \alpha_2,$ $p_2^* = \alpha_1 + (p_1 p_2)\ln\left(\frac{\beta_1}{\alpha_1}\right) + \beta_1 - p_1 p_2, \beta_1 \geq \alpha_2.$

参考文献

[1] Pocock, Stuart J. "Group sequential methods in the design and analysis of clinical trials." Biometrika 64.2 (1977): 191-199.

[2] O'Brien, Peter C., and Thomas R. Fleming. "A multiple testing procedure for clinical trials." Biometrics (1979): 549-556.

[3] Wang, Samuel K., and Anastasios A. Tsiatis. "Approximately optimal one-parameter boundaries for group sequential trials." Biometrics (1987): 193-199.

[4] Gordon Lan, K. K., and David L. DeMets. "Discrete sequential boundaries for clinical trials." Biometrika 70.3 (1983): 659-663.

[5] Jennison, Christopher, and Bruce W. Turnbull. Group sequential methods with applications to clinical trials. CRC Press, 1999.

[6] Bauer, Peter, and K. Kohne. "Evaluation of experiments with adaptive interim analyses." Biometrics (1994): 1029-1041.

[7] 平川晃弘, 五所正彦監訳:臨床試験のためのアダプティブデザイン. 朝倉書店, 2018.

[8] Chow, Shein-Chung. Innovative Methods for Rare Disease Drug Development. 185-204. CRC Press, 2020.

2. Lung-MAP 試験，及び ISPY-2 試験

Lung-MAP 試験[1, 2]は，マスタープロトコルを用いた進行期非小細胞肺がん（扁平上皮癌）の患者を対象とする第 II/III 相試験であり，遺伝子に基づく複数のサブスタディをひとつのマスタープロトコルの下で実施し，各標的治療の安全性や有用性のエビデンスを提供する。被験者はバイオマーカーの結果に基づき 5 つのサブスタディのいずれかに組入れられ，それぞれのサブスタディで試験治療か対照治療が割り付けられる。各サブスタディはそれぞれが第 II/III 相シームレスデザインを用いており，第 II 相試験パートについては奏効率（overall response rate ; ORR）を主要評価項目とする単群試験である。(1) 閾値 15%として奏効率がそれを超える十分な確証が得られること，加えて，(2) 第 III 相試験の症例登録期間がある程度現実的である場合（例えば，想定される症例登録期間が 3 年以下）に限り，第 III 相試験が計画・実施される。第 III 相試験での対照群はその時点で決定される。なお，いずれのバイオマーカーでも陰性であった患者は no-match subgroup に組入れられ，ニボルマブ単独治療とニボルマブとイピリブマブの併用治療の比較が行われる。

ISPY-2 試験[2, 3 ,4]は，早期の乳がん患者において，バイオマーカーで特定したサブタイプに対するネオアジュバンド療法としての新規治療の評価を行う探索的試験である。この試験ではレスポンスアダプティブランダム化（3.2 節）が採用されており，被験者の遺伝的な乳がんサブグループ（3 つの遺伝子マーカーによって 8 つの部分集団に分けられている）において最も効果が期待される治療，または併用治療が割り当てられる。なお，レスポンスアダプティブランダム化では，主要評価項目（早期の代替評価項目）である病理学的完全奏効（pathological complete response; pCR）を用いている。

マスタープロトコルを用いた試験における解析手法については，その目的や試験デザインに応じた適切な手法を計画・適応する必要がある。例えば上述の ISPY-2 試験ではベイズ流階層型モデルを用いた効果の推定を行っている。以下で簡単に触れたい。

ISPY-2 では，バイオマーカーによる病理学的奏効と recurrence-free survival との関係性や予測性を検討するために別途実施された IPSY-1 [5, 6]と ISPY-2 のデータから事後分布を求め，ベイズ流階層モデルに基づく推定値や 95%信用区間の算出や，予測確率に基づくベイズ流の意思決定方法を活用している[3, 4]。ここでは主に Park (2016)の報告に基づき，主要評価項目である病理学的完全奏効の割合の推定方法について以下に概略を紹介する。なお，表記については以下の通りである。

T ： 試験治療 $(T = 0,1,\cdots,\mathrm{M})$，0 は対照群

J : MRI の測定回数，

K: バイオマーカーの数

Y': 開始 6 か月時点における pCR

$Y_j, j = 1,2,\cdots,J$: MRI による機能性腫瘍量（MRI functional tumor volume），連続変数

$Z_k, k = 1, 2, \cdots, K$: バイオマーカーの結果変数, 2 値変数
また, MRI の測定ベクトル, バイオマーカーの結果ベクトルを$\mathbf{Y} = (Y_1, \cdots, Y_J), \mathbf{Z} = (Z_1, \cdots, Z_K)$とする。

＜尤度関数について＞

実際の尤度の算出では, ISPY-2 試験にほかにも別途実施された ISPY-1 試験から得られたデータを重み付きで用いている。この点は後述するが, ここではまず ISPY-2 試験における尤度の算出を説明する。

開始 6 か月時点における pCR の観測値については, 確率 $\mathrm{prob}(Y' = 1|\boldsymbol{Z}, T) = \pi(\boldsymbol{Z}, T, \boldsymbol{\theta}')$の二項分布 $\mathrm{f}(Y'|\boldsymbol{Z}, T, \boldsymbol{\theta}') = \pi(\boldsymbol{Z}, T, \boldsymbol{\theta}')^{Y'}\left(1 - \pi(\boldsymbol{Z}, T, \boldsymbol{\theta}')\right)^{1-Y'}$ に従うとし, この確率$\pi(\boldsymbol{Z}, T, \boldsymbol{\theta}')$のロジット変換後の値を, 定数項（パラメータ$\beta$）, 治療効果（パラメータ$\tau_T$）, バイオマーカーの結果変数（パラメータ$\boldsymbol{\gamma_k}$）, 治療とバイオマーカーの結果変数との 1 次の交互作用（パラメータ$\delta_{T,k}$）, 2 次の交互作用（パラメータα_{T,k_1,k_2}）, 及び 3 次の交互作用（パラメータζ_{T,k_1,k_2,k_3}）の線形和で表した, 以下の Logistic 回帰分析を用いてモデル化する。

$$\mathrm{logit}\left(\pi(\boldsymbol{Z}, T, \boldsymbol{\theta}')\right) = \eta(\boldsymbol{Z}, T, \boldsymbol{\theta}') = \beta + \tau_T + \sum_{k=1}^{K} \gamma_k Z_k + \sum_{k=1}^{K} \delta_{T,k} Z_k$$

$$+ \sum_{k_1=1}^{K-1} \sum_{k_2=k_1+1}^{K} \left(\alpha_{0,k_1,k_2} Z_{k_1} Z_{k_2} + \alpha_{T,k_1,k_2} Z_{k_1} Z_{k_2}\right)$$

$$+ \sum_{k_1=1}^{K-2} \sum_{k_2=k_1+1}^{K-1} \sum_{k_3=k_2+1}^{K} \left(\zeta_{0,k_1,k_2,k_3} Z_{k_1} Z_{k_2} Z_{k_3} + \zeta_{T,k_1,k_2,k_3} Z_{k_1} Z_{k_2} Z_{k_3}\right)$$

この時, パラメータ$\boldsymbol{\theta}'$ には以下が含まれる。

$\beta, \boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_K), \boldsymbol{\alpha_0} = \left(\alpha_{0,1,2}, \cdots, \alpha_{0,K-1,K}\right), \boldsymbol{\zeta_0} = \left(\zeta_{0,1,2,3}, \cdots, \zeta_{0,K-2,K-1,K}\right),$

$\tau_T, \boldsymbol{\delta_T} = \left(\delta_{T,1}, \cdots, \delta_{T,K}\right), \boldsymbol{\alpha_T} = \left(\alpha_{T,1,2}, \cdots, \alpha_{T,K-1,K}\right), \boldsymbol{\zeta_T} = \left(\zeta_{T,1,2,3}, \cdots, \zeta_{T,K-2,K-1,K}\right)$

簡便のため, 対照群（$T = 0$）においては, $\tau_0 = 0$, $\boldsymbol{\delta_0} = (0, \cdots, 0)$ とすると, 上記のモデルに含まれるパラメータは, $\boldsymbol{\theta}' = (\beta, \boldsymbol{\gamma}, \boldsymbol{\alpha_0}, \boldsymbol{\zeta_0}, \tau_1, \boldsymbol{\delta_1}, \boldsymbol{\alpha_1}, \boldsymbol{\zeta_1}, \cdots, \tau_M, \boldsymbol{\delta_M}, \boldsymbol{\alpha_M}, \boldsymbol{\zeta_M})$ で表される。

本試験ではベースライン時, 及び以降 2, 4 か月後, 及び術前に MRI が観測される。MRI には pCR に関する情報が含まれるため, pCR が欠測である場合は MRI の結果に基づいたデータの補完が可能かもしれない。今, MRI についてのベースラインからの変化 $dY_j = \frac{Y_j - Y_1}{Y_1}$ に対し, 以下の分布を仮定する。

$$dY_j \sim \begin{cases} f_{0,j}\left(dY_j \middle| Y', \boldsymbol{Z}, T, \boldsymbol{\theta_j}\right) = \mathrm{N}\left(\mu_{0,j}, \sigma_{0,j}^2\right), & Y' = 0 \\ f_{1,j}\left(dY_j \middle| Y', \boldsymbol{Z}, T, \boldsymbol{\theta_j}\right) = \mathrm{N}\left(\mu_{1,j}, \sigma_{1,j}^2\right), & Y' = 1 \end{cases}$$

ここでは，正規分布$\mathrm{N}\left(\mu_{0,j}, \sigma_{0,j}^2\right), \mathrm{N}\left(\mu_{1,j}, \sigma_{1,j}^2\right)$パラメータを，$\boldsymbol{\theta}_j = \left(\mu_{0,j}, \mu_{1,j}, \sigma_{0,j}^2, \sigma_{1,j}^2\right)$で表した。

今，ISPY-2 試験からn_2例のデータ$\mathcal{X}_2 = \{\boldsymbol{X_i}\}_{i=1,\cdots,n_2}$が得られたとする。なお，症例 i のデータ$\boldsymbol{X_i}$は，MRI の観測値$\boldsymbol{Y_i} = \left(Y_{i,1},,,Y_{i,J}\right)$，バイオマーカーの結果$\boldsymbol{Z_i} = \left(Z_{i,1},,,Z_{i,K}\right)$，受けた治療$T_i$，pCR の観測値$Y_i'$ により，$\boldsymbol{X_i} = (\boldsymbol{Y_i}, Y_i', \boldsymbol{Z_i}, T_i)$ と表す。

この時，ISPY-2 試験における尤度関数は

$\mathcal{L}(\mathcal{X}_2) = \prod_{i=1}^{n_2} L(\boldsymbol{X_i})$

$$= \prod_{i=1}^{n_2} \left[f(Y_i' | \boldsymbol{Z_i}, T_i, \boldsymbol{\theta'}) \times \prod_{j=1}^{J} \left\{ f_{0,j}\left(dY_{i,j} \middle| Y', \boldsymbol{Z_i}, T_i, \boldsymbol{\theta_j}\right) \right\}^{I(Y_i'=0)} \left\{ f_{1,j}\left(dY_{i,j} \middle| Y', \boldsymbol{Z_i}, T_i, \boldsymbol{\theta_j}\right) \right\}^{I(Y_i'=1)} \right]$$

なお，$Y_{i,j}$ が欠測である場合は，$f_{0,j}\left(dY_{i,j} \middle| Y', \boldsymbol{Z_i}, T_i, \boldsymbol{\theta_j}\right) = f_{1,j}\left(dY_{i,j} \middle| Y', \boldsymbol{Z_i}, T_i, \boldsymbol{\theta_j}\right) = 1$ として計算する。また，Y' が欠測の場合は，MRI の直近の測定値を用いて MCMC にて値を補完し，尤度を計算する。

＜事前分布について＞

各パラメータにおける事前分布について，以下のように設定する。

$\beta, \tau_1, \cdots, \tau_M$，及び$\boldsymbol{\gamma}, \boldsymbol{\delta_1}, \cdots, \boldsymbol{\delta_M}, \boldsymbol{\alpha_0}, \boldsymbol{\zeta_0}$ の各ベクトル成分については，すべて独立かつ正規分布 $N(0,5)$ に従う。

$\boldsymbol{\alpha_T}, \boldsymbol{\zeta_T}$ の各ベクトル成分については，すべて独立かつ正規分布 $N(0,0.25)$ に従う。

$\mu_{0,1}, \mu_{1,1}, \cdots, \mu_{0,J}, \mu_{1,J}$ については，すべて独立かつ正規分布 $N(0,1)$ に従う。

$\sigma_{0,1}^2, \sigma_{1,1}^2, \cdots, \sigma_{0,J}^2, \sigma_{1,J}^2$ については，すべて独立かつ逆ガンマ分布 $IG(2.1,1.1)$ に従う。

＜事後分布について＞

事後分布の推定にあたっては，ISPY-1 からのデータ$\mathcal{X}_1 = \{\boldsymbol{X_i}\}_{i=1,\cdots,n_1}$も利用可能であるとし，ISPY-1 試験，及び ISPY-2 試験から得られたデータ$\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2\}$ について，最初の$1, \cdots, n_1$ が ISPY-1 試験からのデータ，次の$n_1 + 1, \cdots, n_1 + n_2$ が ISPY-2 試験からのデータと表記する。また，それぞれの試験から得られたデータに対する重みづけとして，べき乗係数ω_i を以下のように定義する。

$$\omega_i = f(x) = \begin{cases} 0.2, & i = 1, \cdots, n_1 \\ 1, & i = n_1 + 1, \cdots, n_1 + n_2 \end{cases}$$

この時，事後分布は以下のようになる。

$posterior\left(\boldsymbol{\theta'}, \boldsymbol{\theta_1}, \cdots, \boldsymbol{\theta_J} \middle| \mathcal{X}\right) \propto \mathcal{L}(\mathcal{X}_1) \times \mathcal{L}(\mathcal{X}_2) \times prior\left(\boldsymbol{\theta'}, \boldsymbol{\theta_1}, \cdots, \boldsymbol{\theta_J}\right)$

$$\propto \prod_{i=1}^{n_1+n_2}\left[\left\{\pi(\boldsymbol{Z_i},T_i,\boldsymbol{\theta}')\sum_{j\in\mathcal{J}_i}\frac{1}{\sigma_{1,j}}\exp\left(-\frac{\left(dY_{ij}-\mu_{1,j}\right)^2}{2\sigma_{1,j}^2}\right)\right\}^{Y_i'}\left\{(1\right.\right.$$

$$\left.\left.-\pi(\boldsymbol{Z_i},T_i,\boldsymbol{\theta}'))\sum_{j\in\mathcal{J}_i}\frac{1}{\sigma_{0,j}}\exp\left(-\frac{\left(dY_{ij}-\mu_{0,j}\right)^2}{2\sigma_{0,j}^2}\right)\right\}^{1-Y_i'}\right]^{\omega_i}$$

$$\times\exp\left(-\frac{\left(\beta-a_\beta\right)^2}{2b_\beta^2}\right)\times\prod_{m=1}^{M}\exp\left(-\frac{\left(\tau_m-a_{\tau_m}\right)^2}{2b_{\tau_m}^2}\right)\times\prod_{k=1}^{K}\exp\left(-\frac{\left(\gamma_k-a_{\gamma_k}\right)^2}{2b_{\gamma_k}^2}\right)$$

$$\times\prod_{m=1}^{M}\prod_{k=1}^{K}\exp\left(-\frac{\left(\delta_{m,k}-a_{\delta_{m,k}}\right)^2}{2b_{\delta_{m,k}}^2}\right)$$

$$\times\prod_{k_1=1}^{K-1}\prod_{k_2=k_1+1}^{K}\left\{\exp\left(-\frac{\left(\alpha_{0,k_1,k_2}-a_{\alpha_{0,k_1,k_2}}\right)^2}{2b_{\alpha_{0,k_1,k_2}}^2}\right)\prod_{m=1}^{M}\exp\left(-\frac{\left(\alpha_{m,k_1,k_2}-a_{\alpha_{m,k_1,k_2}}\right)^2}{2b_{\alpha_{m,k_1,k_2}}^2}\right)\right\}$$

$$\times\prod_{k_1=1}^{K-2}\prod_{k_2=k_1+1}^{K-1}\prod_{k_3=k_2+1}^{K}\left\{\exp\left(-\frac{\left(\zeta_{0,k_1,k_2,k_3}-a_{\zeta_{0,k_1,k_2,k_3}}\right)^2}{2b_{\zeta_{0,k_1,k_2,k_3}}^2}\right)\prod_{m=1}^{M}\exp\left(-\frac{\left(\zeta_{m,k_1,k_2,k_3}-a_{\zeta_{m,k_1,k_2,k_3}}\right)^2}{2b_{\zeta_{m,k_1,k_2,k_3}}^2}\right)\right\}$$

$$\times\prod_{j=1}^{J}\left\{exp\left(-\frac{\left(\mu_{0,j}\right)^2}{2}\right)\right\}\left\{\left(\sigma_{0,j}^2\right)^{-\left(a_{\sigma_{0,j}^2}-1\right)}exp\left(-\frac{b_{\sigma_{0,j}^2}}{\sigma_{0,j}^2}\right)\right\}\left\{exp\left(-\frac{\left(\mu_{1,j}\right)^2}{2}\right)\right\}\left\{\left(\sigma_{1,j}^2\right)^{-\left(a_{\sigma_{1,j}^2}-1\right)}exp\left(-\frac{b_{\sigma_{1,j}^2}}{\sigma_{1,j}^2}\right)\right\}$$

ただし, 前頁で指定した事前分布の通り,

$a_\beta, a_{\tau_1},\cdots,a_{\tau_M},a_{\gamma_1},\cdots,a_{\gamma_K},a_{\delta_{1,1}},\cdots,a_{\delta_{M,K}},a_{\alpha_{0,1,2}},\cdots,a_{\alpha_{0,K-1,K}},a_{\zeta_{0,1,2,3}},\cdots,a_{\zeta_{0,K-2,K-1,K}}=0,$

$b_\beta^2, b_{\tau_1}^2,\cdots,b_{\tau_M}^2,b_{\gamma_1}^2,\cdots,b_{\gamma_K}^2,b_{\delta_{1,1}}^2,\cdots,b_{\delta_{M,K}}^2,b_{\alpha_{0,1,2}}^2,\cdots,b_{\alpha_{0,K-1,K}}^2,b_{\zeta_{0,1,2,3}}^2,\cdots,b_{\zeta_{0,,K-2,K-1,K}}^2=5,$

$a_{\alpha_{1,1,2}},\cdots,a_{\alpha_{M,K-1,K}},a_{\zeta_{1,1,2,3}},\cdots,a_{\zeta_{M,K-2,K-1,K}}=0,$

$b_{\alpha_{1,1,2}}^2,\cdots,b_{\alpha_{M,K-1,K}}^2,b_{\zeta_{1,1,2,3}}^2,\cdots,b_{\zeta_{M,,K-2,K-1,K}}^2=0.25,$

$a_{\sigma_{0,j}^2},a_{\sigma_{1,j}^2}=2.1,$

$b_{\sigma_{0,j}^2},b_{\sigma_{1,j}^2}=1.1,$

とする。

＜各標的グループにおける pCR の推定について＞

各標的グループにおける pCR の推定については, 以下のように行う。バイオマーカーの結果ベクトル $\mathbf{Z}=(Z_1,\cdots,Z_K)$について, 1 つの数値（Cell Index）$r=1+(2^0\times Z_1)+(2^1\times Z_2)+(2^2\times Z_3)+\cdots+(2^{K-1}\times Z_K)\in(1,2,3,4,\cdots,2^K)$に集約できる。HR, HER2, Mamma Print（MP2）の 3 つのバイオマーカーを取り扱う場合, $\mathbf{Z}$ はr によって以下のように表現できる。

Z_1:HR	Z_2:HER2	Z_3:MP2	r
0	0	0	1
0	0	1	5
0	1	0	3
0	1	1	7
1	0	0	2
1	0	1	6
1	1	0	4
1	1	1	8

ここで, Z_1, Z_2, Z_3 では, 1:陽性, 0:陰性とする

各標的グループは, このセルの集合の部分集団として考えることができる。 $\boldsymbol{R} = (R_1, \cdots, R_{2^K})$, R_r はセル r が各標的グループに属している場合に 1, そうでない場合は 0 とする。上記の例で, HER2 陰性かつ MP2 陰性である集団は, セル$r = 1, 2$ を含んでおり, $\boldsymbol{R} = (1,1,0,0,0,0,0,0)$で表現される。

各セルに所属する確率を$\boldsymbol{\phi}_r$, 試験で各セルに所属する症例数をn_r^* とすると, ベクトル$\boldsymbol{n}^* = \left(n_1^*, n_2^*, n_3^*, \cdots, n_{2^K}^*,\right)$ は , パラメータ n_2 と $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3, \cdots, \phi_{2^K})$ を持つ多項分布 $\text{Multinomial}(n_2, \boldsymbol{\phi})$に従う。ただし, $\sum_{r=1}^{2^K} \phi_r = 1, \sum_{r=1}^{2^K} n_r^* = n_2$ 。

$\boldsymbol{\phi}$ について, ヒストリカルデータである ISPY-1 に基づき, 共役事前分布としてディリクレ分布$\boldsymbol{\phi} \sim \text{Dir}\left(\boldsymbol{\alpha}_{\boldsymbol{\phi}}\right), \boldsymbol{\alpha}_{\boldsymbol{\phi}} = \left(\alpha_{\phi,1}, \cdots, \alpha_{\phi,2^K}\right)$ を仮定する。データ$\boldsymbol{n}^*$ により, 事後分布は$\boldsymbol{\phi}|\boldsymbol{n}^* \sim \text{Dir}\left(\boldsymbol{\alpha}_{\boldsymbol{\phi}} + \boldsymbol{n}^*\right)$ となる。

セルrにおける治療 T の pCR の確率$\pi(r, T)$とセルに所属する確率$\boldsymbol{\phi}_r$ について, MCMC によりサンプル$\pi_s(r, T), \phi_{r,s}$を求めると, 標的グループにおける治療 T の pCR の確率$\pi(R, T) = \frac{\sum_{r=1}^{2^K} \phi_{r,s} \times \pi_s(r,T) \times R_r}{\sum_{r=1}^{2^K} \phi_{r,s} \times R_r}$ を得ることができる。

ここでは pCR の推定方法の概略を記載したが, 参照した Park (2016)の論文補遺は ISPY-2 で用いられている予測確率に基づくベイズ流の意思決定方法などについても詳述されている[4]。

参考文献

[1] Steuer, CE1, et al. "Innovative clinical trials: the LUNG‐MAP study." Clinical Pharmacology & Therapeutics 97.5 (2015): 488-491.

[2] Woodcock, Janet, and Lisa M. LaVange. "Master protocols to study multiple therapies, multiple diseases, or both." New England Journal of Medicine 377.1 (2017): 62-70.

[3] Wang, Haiyun, and Douglas Yee. "I-SPY 2: a neoadjuvant adaptive clinical trial designed to improve outcomes in high-risk breast cancer." Current breast cancer reports 11.4 (2019): 303-310.

[4] Park, John W., et al. "Adaptive randomization of neratinib in early breast cancer." New England Journal of Medicine 375.1 (2016): 11-22.

[5] Esserman, Laura J., et al. "Pathologic complete response predicts recurrence-free survival more effectively by cancer subset: results from the I-SPY 1 TRIAL—CALGB 150007/150012, ACRIN 6657." Journal of Clinical Oncology 30.26 (2012): 3242.

[6] Esserman, Laura J., et al. "Chemotherapy response and recurrence-free survival in neoadjuvant breast cancer depends on biomarker profiles: results from the I-SPY 1 TRIAL (CALGB 150007/150012; ACRIN 6657)." Breast cancer research and treatment 132.3 (2012): 1049-1062.

3. N-of-1 デザインの症例数設定

N-of-1 試験の統合においては，必要症例数として，実施試験数（症例数）だけではなく，各試験（被験者）におけるサイクル数の選び方が重要となる。ここでは，N-of-1 試験の統合に必要な症例数設定の方法をいくつか紹介する。

i. 精度に基づく方法

サイクル数 N の N-of-1 試験を M 個（M 例）実施する状況において，「3.6.2.3 iii Mixed Effect of Difference」で述べたモデルを仮定し，治療効果の推定量の精度ω（分散の逆数）は，次のように導出される[1, 2]。

$$\omega = \frac{\mathrm{M}}{(\tau^2 + 2\sigma^2/N)}$$

上記の式に基づき，被験者間分散τ^2と被験者内分散σ^2の特定の条件下で，目標の精度ωを達成する症例数として，サイクル数 N，及び症例数 M の組合せを決定すればよい。
また，上記の式から，次のような特徴が挙げられる[3]。

- 被験者間分散τ^2，及び被験者内分散σ^2を固定した条件下で，症例数 M が増加するほど，また，各被験者のサイクル数 N が増加するほど，精度ωは向上する。
- 精度を考える上で M と N のどちらが相対的に重要であるかは，τ^2とσ^2の相対的な大きさに依存する。
- τ^2がσ^2に比して小さければ，N を増やすことは有益である。
- 逆に，σ^2がτ^2に比して小さければ，N を増やすよりも，M をより多く増やす方が有益である。

症例数とサイクル数のトレードオフに関して，両者が精度へ与えるインパクトが等価でないことは数値例で示されている[1]。同文献にて，精度を確保するのに膨大な測定が必要なパターンも示されていることから，解析上の効率面だけではなく，症例登録や参加継続といった実施可能性の面においても，症例数とサイクル数のトレードオフを検討する必要がある。詳細については同文献を参照いただきたい。

ii. 検出力に基づく方法

上記の症例数に関する議論は，検出力を考慮して治療間の比較を行う N-of-1 試験のメタ・アナリシスへと拡張できる。例えば，Nikles et al. (2015)[3]は，「3.6.2.3 i Summary Fixed and Random Effect model」を仮定したシミュレーションによる症例数設定を解説している。具体的には，神経性障害性疼痛に対する Gapapentin のプラセボ対照優越性 N-of-1 試験の統合[4]の情報から対立仮説の期待値を設定し，各被験者のサイクル数を 3 に固定した上で，目標の検出力に必要な症例数を計算している。Yelland et al. (2009)[4]では全 3 サイクルを実際に完了した被験者は 48 名であっ

たのに対し，Nikles et al. (2015)[3]のシミュレーションの結果はより少なく，検出力 80%に対して 22 名，90%に対して 33 名という現実的な人数が示された。また，N-of-1 試験の特徴として，多重クロスオーバーにより 1 人 1 人の試験期間が長くなるため，サイクルを完了せずに欠測が生じる問題が挙げられる。この点については，サイクル数の大きさと，少なくとも 1 つのサイクルを完了する被験者の割合や，規定の N サイクルを全て完了する被験者の割合のバランスを考慮し，シミュレーションにより検出力への影響を評価する対応が求められる。

iii. ベイズ流の解析に基づく方法

階層ベイズモデルによる N-of-1 試験の統合においては，公式に基づく方法論が存在しないため，シミュレーションにより症例数設定を行う[5]。ベイズ流の解析に基づくシミュレーションの詳細については，WinBUGS のコードを含む実例[5, 6]が公開されているため参照いただきたい。

参考文献

[1] Zucker, Deborah R., Robin Ruthazer, and Christopher H. Schmid. "Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations." Journal of clinical epidemiology 63.12 (2010): 1312-1323.

[2] Senn, Stephen. "Sample size considerations for n-of-1 trials." Statistical methods in medical research 28.2 (2019): 372-383.

[3] Nikles, Jane, and Geoffrey Mitchell, eds. The essential guide to N-of-1 trials in health. New York, NY, USA:: Springer, 2015.

[4] Yelland, Michael J., et al. "N-of-1 randomized trials to assess the efficacy of gabapentin for chronic neuropathic pain." Pain Medicine 10.4 (2009): 754-761.

[5] Stunnenberg, Bas C., et al. "Combined N-of-1 trials to investigate mexiletine in non-dystrophic myotonia using a Bayesian approach; study rationale and protocol." BMC neurology 15.1 (2015): 1-10.

[6] Stunnenberg, Bas C., et al. "Effect of mexiletine on muscle stiffness in patients with nondystrophic myotonia evaluated using aggregated N-of-1 trials." Jama 320.22 (2018): 2344-2353.

4. Complete N-of-1 デザインの症例数設定

一般化した$K \times J$ クロスオーバーデザインについて，治療効果の推定量$\tilde{D}$に基づく同等性検定における症例数設定の方法[1, 2]を紹介する。試験治療と対照治療の母平均がそれぞれμ_T，及びμ_Rであり，μ_Tのμ_Rに対する同等性マージンを±20%に設定する状況を考える。∇= 0.2とすると，前述の同等性マージンθは$\nabla\mu_R$と表現できることから，同等性仮説は次のようになる。

$$H_0: \mu_T - \mu_R < -\nabla\mu_R \; or \; \mu_T - \mu_R > \nabla\mu_R \; versus \; H_1: -\nabla\mu_R \leq \mu_T - \mu_R \leq \nabla\mu_R$$

このとき，検出力関数は以下の式で与えられる[1, 2]。

$$P(R) = F_\upsilon\left(\left[\frac{\nabla - \mathrm{R}}{CV\sqrt{b/n}}\right] - t(\alpha, \upsilon)\right) - F_\upsilon\left(t(\alpha, \upsilon) - \left[\frac{\nabla + \mathrm{R}}{CV\sqrt{b/n}}\right]\right)$$

各パラメータの意味は以下の通りである。

- R=$(\mu_T - \mu_R)/\mu_R$：相対的な変化量の期待値
- CV=S/μ_R：S はσ_eの推定値であり，各クロスオーバーデザインの分散分析表における MSE の平方根により推定する。
- $t(\alpha, \upsilon)$：自由度υの t 分布の上側 100 α%点。16 順序群，4 ピリオドの 16×4 Complete n-of-1 デザインの場合は$\upsilon = 16n - 5$
- F_υ：自由度υの t 分布の累積分布関数。16×4 Complete n-of-1 デザインの場合は$\upsilon = 16n - 5$
- b：各クロスオーバーデザインに依存する定数。16×4 Complete n-of-1 デザインの場合はb=1/11（本タスクフォースにて導出）。

検出力$1 - \beta$を確保するのに必要な順序群当たりの症例数 n について，$R = 0$の場合は正確な公式が，$R \neq 0$の場合は近似を用いた公式が以下のように与えられる[1, 2]。

$$n \geq b\left[t(\alpha, \upsilon) + t\left(\frac{\beta}{2}, \upsilon\right)\right]^2 \left[\frac{CV}{\nabla}\right]^2 \quad (R = 0)$$

$$n \geq b[t(\alpha, \upsilon) + t(\beta, \upsilon)]^2 \left[\frac{CV}{\nabla - \mathrm{R}}\right]^2 \quad (R \neq 0)$$

数値例を含めた症例数の詳細については，（Chow and Chang 2019，Chow 2020）[1] [2]を参照されたい。

参考文献

[1] Chow, Shein-Chung, and Yu-Wei Chang. "Statistical considerations for rare diseases drug development." Journal of biopharmaceutical statistics 29.5 (2019): 874-886.

[2] Chow, Shein-Chung. Innovative Methods for Rare Disease Drug Development. 185-204. CRC Press, 2020.

2022 年 12 月

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

2021 年度タスクフォース 3 / 2022 年度継続タスクフォース 3

Rare disease の治療効果の推測法

阿多 晃平　　（KM バイオロジクス株式会社）

五十川 直樹　（ユーシービージャパン株式会社）

岡村 正太　　（キッセイ薬品工業株式会社）

小栗 知世　　（アストラゼネカ株式会社（～2022 年 11 月），
ファイザー株式会社（2022 年 12 月～））

島内 順一郎**（アムジェン株式会社）

菅波 秀規****（興和株式会社）

豊泉 樹一郎*（ヤンセンファーマ株式会社）

中田 怜子　　（杏林製薬株式会社）

町田 光陽***（塩野義製薬株式会社）

松尾 一隆　　（ノバルティス ファーマ株式会社）

松嶋 優貴　　（大塚製薬株式会社）

松田 裕也　　（中外製薬株式会社）

森本 賢策　　（日本新薬株式会社）

渡邊 大丞***（サノフィ株式会社）

* チームリーダー，** サブチームリーダー，

*** タスクフォース推進委員，**** データサイエンス部会副部会長