

拘束力のない推奨事項を記載

- 複数の主要有効性（又は安全性）評価項目に対し、各評価項目の役割と重要性の比較を説明することが推奨されている。適切に作成された治験実施計画書は、複数の主要評価項目のそれぞれに対する試験の可否基準を念入りに規定し、多重性の問題を処理し、全体的な第一種過誤率をコントロールする適切な統計的アプローチをあらかじめ決めておくはずである。別の使用目的や強調表示を想定して複数の副次評価項目を選択する場合、データを解析し結果を解釈するための適切な統計方法を治験実施計画書にあらかじめ規定する。
- 代理評価項目は、それがバリデートされており、臨床ベネフィットと直接的な相関関係があれば、使用してもよい。代理評価項目が有効であることが立証されているかどうかを判断するために、そうでない場合は、代理評価項目をバリデートする容認できる方法を決定するために、試験デザインの段階で早期にFDAと話し合うことが重要である。

臨床転帰試験の主要評価項目を選択する際は、以下の事項を検討すべきである：

- 評価項目、転帰、測定値を慎重に選択し、それらが規定されていない、あるいは被験者の相当の割合に対してそれらを取得できないかもしれないという状況は避けるべきである。
- 治験依頼者は、試験をデザインする際に、全体の試験期間（登録完了までの時間とフォローアップ全体の期間を含む）、安全性と有効性の評価項目を評価する時点、インフォームド・コンセントプロセスで被験者がどれくらいの期間フォローアップに同意するかについて、慎重に検討する。特に考慮すべき点は、リスク／ベネフィット解析を行うために安全性と有効性を評価する最初の時点、さらに、承認後に（PMA承認の条件として）フォローアップの追加年数が必要になる可能性である。主要評価項目の時点の選択は、過去の試験のエビデンスを鑑みて、その製品が稼働する時間経過を考慮に入れるべきである。治験依頼者は、承認済みの治験実施計画書とインフォームド・コンセントによって一定の期間のフォローアップが確定された場合、その治験実施計画書を遵守し、登録した被験者をその期間中追跡しなければならない。
- 特定の機器の試験期間中に、科学や医学の解釈が変わった場合、特定の評価項目、転帰、測定値の意義が変わる可能性がある。治験中に試験の評価項目が変わると、特にピボタル試験の場合は、FDAの審査チームと治験依頼者の両方にとって、治験の解釈が難しくなる。適切な対照を設けた治験デザインは、主要要素が全てあらかじめ規定されており、試験の評価項目を含めた主要要素の変更は、治験の解釈とデータ解析に重大な影響を及ぼす可能性がある。このような場合、治験依頼者には、適切なFDA審査部門に連絡し、できる限り最善の措置を話し合うことを勧める。

7.2 臨床転帰試験のインターベンションの割付（無作為化）

21 CFR 860.7(f)(1)に、試験プランや治験実施計画書に以下のことを含めなければならない、と記載されている：

(ii) 以下のような被験者の選択方法：

- (a) 被験者が試験の目的に適していることを十分保証している、治療や診断を行う疾患の診断基準が示されている、場合によっては確認の臨床検査を行う、疾患や症状を予防する機器の場合は、感受性ならびに予防法が必要な症状への曝露のエビデンスを示す；
- (b) 起こりうるバイアスを最小限にする方法で、被験者を試験群に割り当てる（使用する

る場合) ;

(c) 性別、疾患の重症度や罹患期間、被験機器以外の治療法の使用のような、被験群と対照群の関連変数の同等性を確実にする ;

インターベンションや検査の前のベースライン時に両群が同等であり、適切な比較ができるように、被験者をインターベンション群に無作為化することを勧める。無作為化は、性別その他の背景因子変数、疾患の重症度や罹患期間、治療歴、専門ユーザーのバイアス及び／又は好み、治験機器以外のインターベンションの使用のような関連変数に関し、インターベンション群間のバランスが保証されやすい。無作為化は同様に、未測定や未知の共変量を均衡させることも非常に重要である。個々の水準ではなく、異なる水準で無作為化を検討できる場合がある。例えば、治験実施施設の水準では、治験実施施設をインターベンションに無作為化し、その後特定の実施施設の被験者全員を同じグループに割り付ける。これをクラスター無作為化という。治験依頼者は、クラスター無作為化でインターベンションを割り付ける前に、FDAと相談することが望ましい。

並行群間比較臨床転帰デザインの場合、無作為化は通常、各被験者をバイアスのない方法でインターベンションに割り付けるために使用される。各被験者が自身の対照となる一対比較臨床転帰デザインの場合、特定の被験者に使用する2つのインターベンションやインターベンションの部位（例えば顔の右側と左側、左と右の膝）の順番の割付に無作為化を使用すると、バイアスを最小限に抑えることができる。クロスオーバーデザインでは、各被験者に対するインターベンションの順番は概して無作為に決定される。並行群間試験、一対比較試験、クロスオーバーデザイン試験で無作為化しないと、バイアスによって結果がゆがめられ、試験が不成功に終わるリスクがある。

機器のデザインや対象被験者集団が原因でインターベンションの割付を無作為化できない場合、試験に大きさや方向のわからないバイアスが生じ、そのバイアスが試験によって示されたエビデンスの水準と、データが妥当であるとする信頼性に悪影響を及ぼすおそれがある。

当局は、機器の試験で無作為化が不可能、困難又は不適切な状況があることを認識している。例えば、治験責任医師が試験の異なるインターベンションが等しく安全で有効でない（すなわち臨床的均衡がない）と思っている場合、無作為化試験を被験者に勧める際に倫理的ジレンマに陥るかもしれない。そのような場合、治験依頼者は市販前承認の申請や届出を提出する前にFDAに連絡し、無作為化に関する懸念を話し合い、そのような場合でも十分な水準のエビデンスを提供できる適切な試験デザインを決定することが望ましい。

7.3 盲検化（マスキング）

被験者の治療や試験の目的を損なうことなく、インターベンションの割付の情報を制限することを、盲検化、又はマスキングと言う。（マスキングという用語は眼科用製品に適している。）臨床転帰試験という状況で、インターベンションの割付を知っていると、認識しているか否かにかかわらず、被験者、臨床医、治験責任医師、介護者、第三者評価者の行動や決定に影響を及ぼす可能性がある。

被験者のインターベンションへの割付が盲検化（マスキング）されないと、被験者の行動がインターベンションの情報に影響され、その結果、特に臨床測定値や評価項目が主観的な場合、バイアスが混入するおそれがある。

インターベンションの割付が治験責任医師や第三者評価者に盲検化（マスキング）されていないと、治験責任医師や評価者のバイアスが臨床転帰の解釈、機器の外科的埋め込みの性能、その後の臨床決定に影響を及ぼすことによって、試験に悪影響を及ぼすおそれがある。

拘束力のない推奨事項を記載

被験者及び／又は治験責任医師の盲検化が不可能な場合でも、臨床測定値及び／又は評価項目の独立した第三者の評価者をインターベンションの割付に対して盲検化することは可能であり、これが強く推奨される。試験の目的を知らないが、（例えば臨床的、放射線学的な）客観的基準に基づいて評価を行うよう評価者に依頼することが望ましい。あるいは、あらかじめ定めた主な定義と標準操作手順を用いる独立した中央検査機関、測定センター（reading centers）、及び／又は臨床事象委員会を使用し、インターベンションの割付の情報によって評価に影響が及んだ場合に起こるかもしれないバイアスを、最小限に抑えることができる。

機器の臨床転帰試験の中でも、特に侵襲性の高いものや、機器による治療を薬物治療や外科的インターベンションと比較するものは、被験者や治験責任医師をインターベンションの割付に盲検化することができない。しかし、それが不便で難しくても、FDAはできる限り盲検化を検討し、試みることを勧めている。試験を盲検化する場合、試験の最後にインターベンションの割付を盲検化した参加者（被験者及び／又は治験責任医師）に、自分がどちらのインターベンション群に入っていたと思うか、そしてその理由は何かを聞き、試験デザインに盲検化の整合性と有効性の評価を入れると非常に有益であることが多々ある。特に割り付けられた群を知っていると転帰全体に影響する可能性がある場合は、これが試験の結果をさらによく理解するのに役立つと思われる。

試験期間中参加者のインターベンション割付を盲検化しておくことが不可能な場合、できる限りバイアスを最小限に抑えるために以下の方法を検討する：

- 被験者候補をスクリーニングし、登録が終了する直前まで、被験者と試験スタッフにインターベンション割付を盲検化する。
- インターベンション割付を知ったことによる試験脱落の事態を避けるために、処置が終わるまで被験者を盲検化することを強く推奨する。
- 盲検化されない場合、被験者報告転帰よりもさらに客観的な評価項目が望ましい。被験者報告転帰を採用する場合、試験に関連する臨床スタッフが立ち会わずに測定されるように考慮する。
- 試験参加者にする追跡調査の質問を標準化するために、臨床スタッフが使用するスクリプトの草稿を作る。

7.4 比較臨床転帰試験の対照

21 CFR 860.7(f)(1)(iv) に対照の4つの種類が記載されている。試験プランや治験実施計画書に、以下のことを含めなければならない、と記載されている：

量的評価ができるような方法での治療結果や診断結果の対照との比較。対照の正確な性質を明記し、観察者とデータ解析者のバイアスを最小限にするために採用した方法を説明する。「盲検化」の水準と方法が、適切で使用する場合は文書化する。一般に、4つのタイプの比較が認められている；

(a) 無治療。有効性のより客観的な測定値があり、プラセボ効果を見逃せる場合、治療した患者と未治療の患者の比較群で客観的結果を比較する；

(b) プラセボ対照。機器の使用によるプラセボ効果があるかもしれない場合、できるだけ離れた試験で、かつ使用条件と類似させた条件下で、その機器と無効の機器の使用結果を比較する；

拘束力のない推奨事項を記載

(c) 実治療対照。例えば治療する疾患が、プラセボの使用や治療の中断が不適切であり、患者の利益に反するような疾患であり、比較に有効な治療レジメンを使用できる場合；

(d) 歴史的対照。ある特定の状況で、死亡率が高く、死亡が予測できる疾患を有する患者、又は罹患期間や重症度が予測できる徴候や症状がある患者、又は罹患が予測できる予防法の場合、機器の使用の結果を、治療を受けなかったか確立されている有効な（治療、診断、予防の）レジメンに従った同等の患者や集団の疾患や症状の、適切に文書化された自然な疾患経過に由来する過去の使用経験と量的に比較できる。

CFRに記載されている4つの種類の対照に加えて、本指針は、5番目の「被験者自身対照」も検討する。本指針では、上記の(a)と(c)で対照を説明する際に、「治療」の代わりに「インターベンション」という用語を使用している。この用語は診断インターベンション、さらに治療インターベンションと美容インターベンションの臨床転帰試験に当てはまるからである。

臨床試験に使用する場合、それぞれの対照にメリットと限界がある。一般的に、非同時的対照を用いた試験デザインより、同時的対照を用いた試験デザインの方が、バイアスは少ない。

表1に、各対照と試験のバイアスの関係、それによるエビデンスの水準について説明する。

表1：臨床転帰試験の対照の種類

対照の種類	サブカテゴリー	説明	考察点
同時的対照	実インターベンション対照（「実対照」）	対照群は、既知の効果をもたらす別のインターベンションを行う（通常は別の機器や手術、ただし薬物や生物製剤の可能性もある）	<ul style="list-style-type: none"> ● 実対照に対する優位性又は非劣性を実証する ● 適切な対照の選択は、標的被験者集団に対する現在の標準治療に基づいて行う ● 実対照の効果の大きさについてどれだけわかっているか
	プラセボ対照（「みせかけ」）	対照群は治療（又は診断）効果がないと考えられている別の機器、シミュレーション処置又は薬物や生物製剤。	<ul style="list-style-type: none"> ● プラセボ対照は、プラセボ効果があると考えられている場合に有用である。 ● 治験機器のような機能を持っているように見えるが、治療を行わないプラセボ対照を作ることは難しい。 ● 既知の効果がないプラセボに被験者を無作為化することが非倫理の場合もある。
	「インターベンションを行わない」対照	対照群がインターベンション（又は診断）を行わない。	<ul style="list-style-type: none"> ● 「インターベンションを行わない」対照の選択は、被験者募集に問題が生じる可能性がある。被験者がインターベンションを受けられず、「インターベンションを行わない」対照群に無作為化されて登録し続けることになる。

拘束力のない推奨事項を記載

対照の種類	サブカテゴリー	説明	考察点
			<ul style="list-style-type: none"> 対照群の被験者はベネフィットを期待できず、実験群の被験者はベネフィットを期待できるので、「インターベンションを行わない」対照の選択は、それ自体にバイアスがある。 「インターベンションを行わない」対照は、ベネフィットやリスクの漸増のエビデンスを提供できる標準治療／最善の医学的管理であることが多い。ただし、この対照は治験実施施設によって異なる可能性がある。
	被験者自身が対照になる	被験者が自分の同時的対照になる(例えば split face、fellow eye など)。	<ul style="list-style-type: none"> 治療及び／又は美容機器の試験の場合、被験者自身を同時的対照に使用すると、その被験者における相関関係を利用できる。 このデザインは、実験機器と対照インターベンションの効果が局所的で、重複しない場合にのみ可能である。
非同時的対照	被験者自身が対照になる	被験者のベースライン時の転帰を、エンドポイント評価時の転帰と比較する。	<ul style="list-style-type: none"> エンドポイント評価時の転帰の比較にベースライン転帰を使用することは、ほとんどの治療試験において不適切である。被験者が治験機器と関連性のない理由(例えば平均への回帰、プラセボ効果)で改善している可能性がある。
歴史的対照	並行群の被験者水準データ	対照群は、過去に治療を受け、現在の試験と同じ転帰及び同じ共変量に対して個々の被験者水準のデータがある、様々な被験者群で構成される。	<ul style="list-style-type: none"> 重大な問題は、両群のベースライン共変量に関する同等性である。時間を隔てた比較試験を使用すると、重大な未知の選択バイアスが混入するおそれがある。 共変量解析や傾向スコア解析のような統計的手法により対処できる問題もある。 歴史的対照群は現在の医療方法を反映していない可能性があり、現代の試験とは異なる被験者集団及び／又は転帰が入っている可能性がある(時間的バイアス)。 主観的評価項目が使用されているときや、必要な評価項目が全てこれまで評価されなかったことがないか別の方法で評価されている場合、この対照は難しい。 この対照は、欠損データの推測が非常に難しい。感受性と欠損データの解析によって、バイアスに関する問題に対処できる場合もある。

7.5 プラセボ効果とその他の現象

多くの臨床転帰試験の問題点は、機器に実際の効果はないかもしれないが、それでも有効性が実証されるように見えるという点である。これに対抗するために、プラセボ機器(「見せかけ」機器と言うこともある)を使用する。プラセボ機器は、明らかな効果をもたらさないように意図的

拘束力のない推奨事項を記載

にデザインされているが、それでも有効性を実証するように見える。この現象をプラセボ効果という。プラセボ効果は、疼痛、機能、生活の質の試験でよく起こり、非常に大きくなる可能性もある。プラセボ効果は客観的評価項目と主観的評価項目に観察され、何ヶ月、何年も持続することがわかっている。

プラセボ効果がよく認識されている理由はいくつかある。

- **ベネフィットの期待**—無作為化盲検化試験で、被験者はどちらかの群に無作為化されるのでベネフィットの期待がある；これは、「インターベンションを行わない」対照群がある、被験者がベネフィットを期待できない非盲検化試験とは対照的である。
- **試験効果**—プラセボ効果に関連するのは、人は、試験で測定されていることがわかると行動が変わる傾向にあるという点である。さらに、被験者は、試験でより手厚いケアを受ける。これらの効果の両方が、どの試験でも客観的及び主観的に報告される転帰に影響を及ぼす可能性がある。

プラセボ効果によって、治験機器と対照機器による改善の単純な比較にバイアスが混入する。このため、治験機器と有効性のない治療法を比較することが可能な場合は、プラセボ対照を使用することが望ましい。プラセボに対する優位性が実証できれば、治験機器が有効であることが推定できる。このような試験は、インターベンションの割り付けが被験者、治験責任医師、第三者評価者に盲検化されているときに最も機能する。

実対照を使用するとプラセボ効果を直接測定できないが、両インターベンション群でプラセボ効果が同等であると想定できる場合、両群の相対的な安全性と有効性を適切に比較できる。残念ながら、実対照のある無作為化試験では、各群のプラセボ効果の大きさが異なる可能性があり、それは試験手順に関連する「形式」とほぼ比例している。（例えば、観血的手術は経口薬の服用よりプラセボ効果が大きい）。

診断臨床転帰試験において、様々な治験実施施設の臨床医に異なる標準治療がある場合、試験結果の解釈に問題が生じる可能性がある。プラセボ対照試験に最も近いものは、標準治療に関する行為の変化を最小限に抑えるために、機器を使用した後まで自分の被験者がどの群にいるかを臨床医が知らない試験かもしれない。

これに関し、試験結果の解釈を難しくするその他の現象がある。

- **平均への回帰**—無作為の要素がある被験者の測定値の場合（例えば試験の選択基準で定められているような）試験の登録時にその被験者に極端な測定値がある場合、同じ被験者のその後の測定値は全体の平均に近づく傾向がある。よって、純粹に統計的な現象として、最初に「具合が悪い」被験者ほど改善が大きくなる傾向がある。
- 臨床試験では医学的配慮が大きい被験者が改善に至る可能性がある。
- 自発的寛解—試験中に自然に治癒する、あるいは症状が出なくなる被験者もいる。

7.6 非比較臨床転帰試験

臨床転帰試験デザインの中には、同時的（又は歴史的）対照を用いていないことから直接の比較対照がないため、適切な対照を設けていない試験がある。

7.6.1 客観的性能基準（OPC）による単一群試験

客観的性能基準（OPC）とは、臨床試験及び／又はレジストリの歴史的データから導き出した標的数値を指し、FDAが安全性や有効性の評価項目の審査や比較での二分法（合格／不合格）に用いることがある。現在バリデートされているOPCは非常に少ないことを言う必要がある。OPCは通常、機器のテクノロジーが十分成熟しており、公知の情報や特定の種類の機器に関するあらゆる試験からプールされた情報に基づいている場合に、作成される。OPCは、重要な全ての情報源の過去のメタアナリシスレビューから慎重に構築する必要があり、被験者水準のメタアナリシスが望ましい。OPCは、医学学会や科学学会、標準化機構によって委託された、又は採択された場合、あるいはFDAの指針文書に説明されている場合、妥当性が大きい傾向がある。OPCは一般的に、1つの企業がそのデータや医学文献のレビューに基づいて作成できるものではないし、FDAが一方的に作成するものでもない。また、OPCはテクノロジーが成熟し、改良されるにつれて廃れるということにも留意しなければならない。ピボタル臨床試験にOPCを用いようとする治験依頼者は、試験開始前にFDAスタッフと話し合い、同意を得るべきである。

7.6.2 性能目標（PG）による単一群試験

性能目標（PG）のエビデンス水準は、OPCより劣っている。PGは、FDAが、安全性及び／又は有効性の評価項目の比較として使用に十分であると見なした数値（点推定値）を指す。PGが、有効性及び／又は安全性評価項目の信頼区間の上限（又は下限）限界値に基づいている場合もある。一般的に、機器のテクノロジーは、OPCほどPGの使用に対して十分開発されていない又は成熟していないので、PGを作成するためのデータは、OPCの作成に使用できるデータほど堅牢と見なされていない。PGは、問題のある患者集団や対照に臨床的釣合いがない場合に検討できるかもしれない。PGは、医学学会や科学学会、標準化機構によって容認、又は作成された場合、あるいはFDAの指針文書に説明されている場合は、妥当性が大きい傾向がある。治験依頼者がPGを創作することや、FDAが特定の申請のために一方的に作成することは通常推奨されていない。OPCと同様、PGも、テクノロジーが改良され、機器の性能についての情報が増えるにつれて廃れる可能性がある。ピボタル臨床試験にPGを用いようとする治験依頼者は、試験開始前にFDAスタッフと話し合い、同意を得るべきである。

PGは非常に注意して使用する必要がある。特に、重要な問題は、安全性（又は有効性）の性能目標に達した機器が実際に、適切な対照を設けた試験でその安全性（又は有効性）を実証できるという確実なエビデンスがあるかどうかということである。PGに達した（あるいは達しなかった）からといって、必ずしもその試験の結果を直ちに容認できる（またはできない）ということにはならない。試験の結果を併合する場合や、その結果に通常ではない兆候が見つかった場合には、試験結果の質をさらに探索する必要がある。FDAは、PGを用いたPMAを関連諮問委員会に提出し、試験結果の解釈について外部の科学的意見を得ることがある。

7.6.3 観察試験又はレジストリ

治療効果を比較するために臨床データベースを検証することは、バイアスの危険をはらんでいる。臨床試験における無作為化は予後に基づく治療の割付を防ぐが、臨床転帰のある観察試験及びレジストリにおいてこの種のバイアスがコントロールされるという保証はない。無作為化臨床試験の転帰が観察試験で報告されているものと大きく異なるという文献の例がある。この不一致の1つの理由は、観察試験のインターベンションの割付が被験者の予後によって左右されたかもしれないということである。

疫学的研究に使用される他のデザインには、マッチする重要な共変量に基づいて選択された1つ又は複数の対象被験者の症例が必要かもしれない。選択される症例が標的集団全体のサブセットから不均等に選ばれる可能性があり、よって対照も標的集団の典型ではない可能性があるため、マッ

拘束力のない推奨事項を記載

チングが問題になると思われる。この種の観察試験は、診断か治療かにかかわらず、市販前試験には推奨されていない。しかし、特定の事象と特定の機器を関連づける市販後試験に有用な場合がある。

メタアナリシス

新しい臨床データを作成せずに、医療機器の安全性と有効性をメタアナリシスで実証すると、バイアスが混入する可能性がある。重要でない結果や悪い転帰の試験は通常発表されないからである。まれにこの試験デザインが有用である場合は、正確な統計的手法を使用し、発表文献の選択基準と却下基準にあらかじめ決定した厳しい品質管理を用いて選択のバイアスを最小限に抑えることが重要である。十分容認されているメタアナリシスの方法論は、解析を始める前に、メタアナリシスに含める試験（例えば無作為化臨床試験）の選択基準を明確にすることである。このアプローチを前向きメタアナリシスという。しかし、重大な欠点は、出版物の大半に被験者水準のデータや、各試験のデータを独自に解析できるような十分に詳しい内容が入っていないことである。その他よくある問題として、選択/除外基準が試験によって異なる、評価項目の定義が大きく異なる、被験者の追跡期間が異なる、などが挙げられる。メタアナリシスには治験依頼者が販売したいバージョンの機器の試験のみを入れるようにすることが重要である。

7.6.4 文献サマリー

文献サマリーには、有資格の専門家が行った適切な証拠文書のあるケースヒストリー、市販機器の重要なヒトの使用経験の報告書などがある。

これらの報告書では、新しい臨床データは作成されていないが、新しい解析が行われないメタアナリシスとは異なる。文献サマリーが含まれているPMAは、選択された発表済み文献で行われた解析、さらにおそらく特定の治験責任医師の適切な証拠文書のある経験に頼って審査される。これらの報告書は、メタアナリシスの使用よりはるかに多くの重大な制限があるので、有効性の実証にはほとんど有用ではない。

7.7 診断臨床転帰試験

診断機器の場合、ピボタル臨床試験は診断臨床性能試験であることが多い（第8項参照）。しかし、臨床転帰試験が必要な場合もある。診断臨床転帰試験では、治療や管理のインターベンションに診断機器の結果を使用する。インターベンションが被験者の転帰に与える影響によって、機器の性能の一部が評価される。診断臨床性能も評価できる。臨床転帰試験は、例えば、疾患や症状の診断と治療が同時に行われる場合（例えば一部の内視鏡処置）や、正確な診断による臨床ベネフィット（臨床転帰の改善）が明確でない場合に、適切かもしれない。検体を採取するためだけに必要で、試験でその診断結果を管理方法の決定に使用しないインターベンションは、本指針では診断臨床転帰試験と見なされない。

適切な臨床評価項目か診断性能、又は両方で、安全性と有効性が測定される。安全性と有効性を十分評価するために、臨床医が診断結果を用いない対照群が必要な場合がある。治験群と対照群の比較には、並行群や一対比較デザインが適切かもしれない。

対照を設けた診断臨床転帰試験には、臨床医自身が診断結果を使用しているかいないか知っているため、被験者が治験群にいるか対照群にいるかを臨床医に盲検化できないものがある。しかし、できる限り、臨床評価項目を評価する臨床医を、被験者のいる群に対して盲検化すべきである。

7.8 臨床転帰試験の長所と短所

所定の機器や目標とする使用目的に適切な試験デザインを決定する場合、機器の特徴、使用条件、同じ使用目的に対する代替インターベンション（又は診断検査法）の有無、機器の使用に関する十分な警告の有無、機器の使用経験の程度など、多くの要因に左右される。試験はラベル表示を裏付ける十分なエビデンスを提供する必要があるため、最終的に目指すラベルの強調表示と使用方法を検討することも重要である。FDAが推奨するのは、治験依頼者が、特定の機器の制約や特徴を踏まえ、その使用目的に対する機器の安全性と有効性を合理的に保証するために必要なエビデンスが提供できそうな試験デザインを選択することである。

他よりも高水準のエビデンスを示すことができる試験デザインもある。低水準のエビデンスを提供する試験デザインを選ぶ場合は、そのデザインが適切であり、潜在的バイアスを十分抑制して、使用目的を裏付けられるという理由が必要である。治験依頼者が臨床転帰試験に適切な対照設定、無作為化、及び／又は盲検化が適切でない、必要でないと考えられる場合、なぜバイアスの可能性を無視できるかを必ず説明しなければならない。バイアスを最小限に抑えるような試験をデザインするほど、（他が同じであれば）エビデンスの水準は強くなる。

次の項では、臨床転帰試験に多く用いられる試験デザインの長所と短所について説明する。

7.8.1 無作為化、二重盲検、対照、並行群間臨床試験

この試験デザインは、科学エビデンスの水準が最も強く、バイアスの量が最も少ないので、概して並行試験デザインを企画する際に望ましい。二重盲検とは、被験者も試験スタッフ（治験責任医師や第三者評価者を含む）もインターベンションの割付を知らないことを指す。この試験デザインは、治験群と対照群の被験者集団が同等であることを保証する水準が最も高く、安全性と有効性の転帰に影響を及ぼす可能性がある既知と未知のベースラインの変数に関し、群間の系統的な差を回避できる。しかし、このデザインが実現できない又は現実的ではない機器がある。

この試験デザインに選択される対照は実対照かプラセボ対照である（7.5項参照）。この試験デザインからの逸脱は、プラセボ効果の可能性がある場合や、試験評価項目として主観的な転帰測定値が使用される場合に特に問題である。このようなデザインは治験機器のベネフィットとリスクの直接的なエビデンスを提供できるので、プラセボ対照の使用が望ましいが、対照群の被験者の治療を阻むことがしばしば問題となる。したがって、実対照やプラセボ対照の選択は倫理面や実際の状況によって左右される可能性がある。実対照を検討する際の重要事項は、実対照と比較する場合に優位性と非劣性のどちらを実証するために試験をデザインするかということである。

7.8.2 被験者自身が対照となる無作為化一対臨床試験

この試験デザインは、被験者に治験インターベンションと対照インターベンションの両方を同時に行う。例えば、顔の半分を治験機器で治療し、他の半分を対照インターベンションで治療する。このデザインでは、インターベンションの割付（顔のどちら側か）が無作為化される。この試験デザインは、機器の局部への効果が見える場合のみ使用できる。この試験デザインでは、全身の安全性や有効性の転帰を評価し、区別することはできない。この試験デザインを適切に使用した場合の長所は、同じ被験者で両インターベンションの効果を測定でき、さらに変動性が小さいので必要とされる標本数が少なくよいことである。

この試験デザインのもう一つの種類は、2群のクロスオーバーデザインである。ここでは各被験者が、無作為に割り付けられた治験インターベンションと対照インターベンションを順番に受ける。このデザインも同様に、各被験者で治験機器と対照インターベンションの性能を比較することができる。しかし、この場合、最初のインターベンションの効果が2番目のインターベンション期間

に持ち越されないことを想定する必要がある。この想定が適切でないときは、インターベンションの間の期間（「ウォッシュアウト期間」）を長くしなければならない。

7.8.3 同時的対照（実対照、プラセボ対照又は「インターベンションなし」）を設けた無作為化、非盲検化試験

同時的対照を設けた無作為化非盲検化試験と、7.8.1及び7.8.2の試験デザインとの主な違いは、盲検化が不完全である、又は盲検化されていないことである。不完全な盲検化とは、被験者、治験責任医師又は第三者評価者が盲検化されていない場合を指す。誰も盲検化されていない試験は、オープンラベル試験ともいう。上述したように、比較臨床試験の場合、被験者、治験責任医師、第三者評価者をインターベンションの割付に盲検化すると、バイアスを最小限に抑えることができる。しかし、実対照やプラセボ対照があると、被験者や治験責任医師を必ずしも盲検化できず、第三者評価者を盲検化することが難しい場合もある（例えば、治験機器と実対照の機器が画像上で全く異なる外観を呈する場合）。

対照が「最善の医学的管理法」や「インターベンションなし」である場合、この試験は通常、被験者にも治験責任医師にも盲検化されない。したがって、対照群の全ての被験者は、自分が治験機器を使用されていないことを知っている。このことによって、未知の大きさのバイアスが生じることがある。

試験参加者が無作為化されていない場合、それによって生じるバイアスの大きさを評価することは非常に難しく、真に客観的な評価項目を使用しても、それ以外は信頼できる試験の科学的妥当性を脅かす可能性がある。試験参加者（被験者、治験責任医師、評価者）のいずれか又は全員の盲検化が不可能な場合、バイアスに関する問題の対応策の根拠と説明を、FDAに詳しく示す必要がある。

7.8.4 同時的対照（実対照、プラセボ対照又は「インターベンションなし」）を設けた非無作為化試験

同時的実対照を設けた非無作為化デザインでは、被験者と治験責任医師がインターベンションの割付に無作為化されない。したがって、この試験デザインには、同時的対照デザインを設けた無作為化非盲検化試験の欠点がある。さらに、無作為化がなく、各被験者が可能なインターベンションのうち1つだけを受けるので、インターベンション割付に起因する未知の大きさのバイアスが非常に現実的な可能性がある。

このデザインは無作為化試験と同じぐらい労力を要するが、おそらく群間差、実施施設、治験責任医師の差、さらに測定はされないが交絡が生じると思われる差により大きなバイアスが混入するので、一般的に推奨されていない。試験全体の2つのインターベンション群が均衡しているように見えても、治験責任医師と機器の相互作用がある（治験機器の長所が治験責任医師によって違って見える）というように、参加している各治験責任医師が均衡していない可能性がある。

7.8.5 ベースラインと比較する単一群試験

多くの治療法の試験で、非常に重要な事項は、対照として被験者のベースライン状態を使用することは魅力的であるが、通常は実対照やプラセボ対照（あるいは「インターベンションなし」の対照でも）を設けた無作為化群も設定するのが望ましい。盲検化試験のこのような無作為化群は、非常に説得力のある対照を設定でき、プラセボ効果によるバイアスや時間的バイアスを回避できる。

7.8.6 歴史的な対照や情報がある単一群試験

機器のテクノロジーが十分開発され、対象の疾患がよくわかっている場合、歴史的対照や何らかの歴史的情報がある単一群試験を行うことができる。

- 被験者水準のデータがある歴史的対照群との比較：

過去の試験と現在の試験の両方に各被験者の重要な変数を全て含めた被験者水準のデータがある場合、少なくとも何らかの統計的比較は可能である。難しいのは、その歴史的対照が現在の試験群と同等であることを実証することである。現在の試験が終わった後、傾向スコアモデルを用いて両群の同等性を比較することができる可能性がある。ただし、最後にデータが同等にならないかもしれないという大きなリスクがある。被験者が登録され、ベースラインデータを集めて解析するまで同等性の評価をする方法がないので、このアプローチはリスクがある。

両群が同時ではないので、歴史的対照の使用に内在する明らかなバイアスは、時間的バイアスである。この時間の隔たりにより2つのインターベンション群の同等性についての問題が生じ、さらに医療方法が変化し、その結果標的被験者集団と予測される転帰も変化するという問題が生じる。したがって、このデザインの短所は、歴史的対照の被験者の転帰を、標的とする現在の集団と識別したり、これに適用したりできないことである。

- 歴史的情報から導き出したOPC やPGとの比較

歴史的対照群がない場合、安全性や有効性の評価項目に関する目標数値であるOPCやPGで機器の性能を評価することができる。このような試験デザインは、歴史的対照との比較と同じ問題点や限界点がある。さらに、そのOPCやPGを導き出した歴史的対照群と現在の群がどれくらい同等かを評価する独立した方法がない。さらに、バイアスの量測定ができない。

このような試験には対照群がないので、OPCやPGとの比較は、優位性や非劣性を実証できない。

7.9 IDE申請に関する事項

臨床転帰試験の場合、21 CFR 812.20に基づいて、治験依頼者のIDE申請に、提案する試験デザインとそのデザインを選んだ理由を詳細に記載しなければならない。治験依頼者は、その試験デザインが何故十分に堅牢なのかを説明しなければならない。検討されたが却下された、バイアスの可能性が少ない別の試験デザインを考察すると、提案する試験デザインの効率的な審査に役立つかもしれない。

8. 診断臨床性能試験

診断機器の場合、ピボタル臨床試験が診断臨床性能試験であることが多い。この試験の場合、診断機器の出力が、被験者の標的疾患の評価に使用される臨床参照基準（8.2項に定義）とどれくらい近い関係にあるかを測定する方法で、診断機器の診断臨床性能を特徴化する。適切な臨床性能測定法の選択は、機器の使用目的、診断機器の出力の性質、臨床参照基準に左右される。診断臨床性能試験の目標は、機器の診断臨床性能を実証し、標的集団における当該機器の臨床性能について良好なリスク／ベネフィット解析を裏付けることである。

診断臨床性能試験の前にベンチ試験、非ピボタル臨床試験及び／又は機器の測定値に関する様々

拘束力のない推奨事項を記載

な側面を分析する試験（例えばIVDの分析試験）が行われることが多い。例えば、ヒト乳頭腫ウイルス（HPV）DNAの高リスク菌株を検出し、Pap検査結果が正常な30歳以上の女性の子宮頸がんを予測する*in vitro* 診断機器を考えてみよう。子宮頸がん（標的疾患）を予測するHPV検査の能力は臨床性能試験で評価されるが、HPV DNAの高リスク菌株（測定対象）を測定する機器の能力は、分析性能試験（一例を挙げると、測定バイアスの評価、精度、定量化と検出の限界、直線性、干渉、キャリーオーバー）で評価される。

診断機器の安全性と有効性は分けられないことが多い。診断機器によって報告された結果が間違っている（例えば、その結果が偽陽性や偽陰性に間違っ分類された）場合や誤って解釈された場合、その後の不適切な管理や心理的なトラウマによって被験者が傷つくおそれがある。標的疾患が二分法（あるかないか）で診断されるものである場合、診断機器の安全性と有効性は、標的疾患があるかないかを正しく識別する能力ととらえられることが多い。誤診断に加え、診断機器は、検体の採取や機器の使用中に被験者に安全性の懸念が生じることがある。例えば、患者を放射線その他のエネルギーに曝露させる、侵襲的な処置を使用する結果になる、治療薬を投与する結果になるというような可能性がある。このような場合、試験デザインの適切さを評価し、その機器が試験の被験者に重大なリスクを与えるかどうかを判断する際に、診断臨床性能試験における被験者へのリスクも考慮しなければならない。

診断臨床性能試験のための臨床試験のデザインに影響を与える重要な要因には、試験デザインを決めるための機器の使用目的の重要性、適切な試験集団の選択、特定のバイアス原因の軽減などがある。下記の項でこれらの要因をさらに詳しく論じる。

8.1 使用目的

診断機器のもたらす結果の種類が多様であるため、これら機器の使用目的は大きく異なる。したがって、診断臨床性能試験のデザインも多様である。多くの診断機器は、標的となる症状、疾患の有無やステージに従って被験者を分類しようとする。診断的評価や被験者のモニタリングの一助として、生物学的量（例えばウイルス負荷、血糖値、網膜の厚さ）の測定値を示す診断機器もある。

ピポタル診断臨床性能試験は診断機器の使用目的を裏付けなければならない。診断機器には、既存の診断機器や診断処置に代わる独立した診断機器として使用するものや、他の情報と組み合わせ（時にはアルゴリズムの使用によって）被験者の標的疾患を評価するものがある。また、診断機器が補助的な診断情報を提供する場合もある（例えば、その追加の情報で既存の機器や処置に基づく推奨方法を裏付ける）。

診断性能試験をデザインする場合、以下のような使用目的に照らして機器を評価すべきである：

- その機器は何を測定／検出するか
- その機器は何を報告するか
- 検査される細胞、組織、臓器、身体の一部又は全身
- 検体の採取元、検体の種類、検体マトリックス
- 機器の使用法（使用説明書に基づく）
- 機器の使用時期（使用条件）
- 機器の使用者（オペレータや標的ユーザー）

拘束力のない推奨事項を記載

- 何を対象とするか（標的疾患）
- 誰を対象として機器を使用するか（標的集団）

既に承認や許可がなされている機器の使用方法を変える場合は、どんな評価が必要かをFDAと相談するとよい。

8.2 標的疾患の臨床参照基準

理想を言うと、診断機器の臨床性能の分析には、標的疾患に関する被験者の真の状況の中立的な情報が必要である（4.1項参照）。診断臨床性能試験において、臨床参照基準を用いて各被験者の標的疾患が評価される。規制上、臨床参照基準は、標的疾患に関する被験者の真の状態の実証に最もよく使用されている方法である（例えば、乳癌の有無を判断する生検の病理結果）。1つの方法である場合もあれば、臨床追跡調査のように方法や手法が組み合わされている場合もあるが、治験機器の出力を考慮してはならない。

別のFDA指針17に説明されているように、「最もよく使用されている方法」は、医学会、検査学会、規制当局内の現行の方法によるエビデンスで確立されている。いくつかの可能な方法から1つが選ばれることもある。複合体が作成されたり、専門家グループのコンセンサスが使用されたりする場合もある。試験を開始する前に、臨床参照基準をあらかじめ決めておくべきである。また、その結果を判断するために必要な情報を含め、十分詳しく説明すべきである。できる限り、この必要な情報を試験の各被験者に対して記録すべきである。各被験者に対して参照基準の結果を判断することが望ましい。

臨床参照基準がない、入手できない、その侵襲性のために臨床試験に使用できない場合がある。そのような場合、臨床参照基準でない対照（非参照基準）を定めて使用することができ、この結果を治験機器の出力と比較する。例えば、被験B型肝炎アッセイは、通常、B型肝炎ウイルスの臨床参照基準ではなく、FDAが承認している複数のHVBマーカーアッセイの結果と比較される。治験依頼者は、臨床参照基準を使用しない試験を計画する前にFDAに相談し、試験が機器の使用目的を裏付けるか確認すべきである。標的疾患の評価に臨床参照基準を使用しない診断臨床性能試験を、*agreement study*（一致試験）という。一致試験では、診断機器の「正確さ」を直接検証できない。治験機器は非参照基準の比較機器と一致するかもしれないが、もし臨床参照基準が使用できたら出たであろう結果と一致しないかもしれないからである。一致する測定値の解釈に関する問題点が、他のFDA指針で2つの診断機器の結果の比較に照らして論じられている。

臨床参照基準は、情報が増え、医療システムが発展するにつれ、時間と共に進化するので、診断臨床性能の測定値を、使用した臨床参照基準とともに報告し、さらにその背景の中で解釈すべきである。検査開発者が、臨床参照基準のエラーによって新しい機器と臨床参照基準の間に不一致が生じると考える場合、治験依頼者は試験実施前にFDAに相談すべきである。

8.3 診断性能の評価のための試験集団

機器の使用目的を裏付ける試験のために、被験者や標本が選択される治験実施施設は、その機器がその目的のために使用される施設でなければならない。被験者や標本も、提案する標的集団の代表者でなければならない。代表者でない施設や被験者から全体の性能を推定すると、選択バイアスが混入する。その試験の被験者や標本を選択する実際の方法を、治験実施計画書に明記しなければならない（21 CFR860.7(f)(1)(a)参照）。色々な選択方法と、その長所及び欠点が前の6.3項に説明されている。

試験に登録された被験者は、標的疾患の範囲にいなければならない。登録した被験者が標的疾患

の範囲に当てはまらないと、診断臨床性能の推定値に範囲の影響が及ぶ。例えば、その標的疾患の両極端にいる被験者（例えば健康で正常な被験者や進行ステージにいる被験者）だけから標本を取ると、性能は実際よりも優れているように見える。抜けた標的疾患の範囲の真ん中にいる被験者の方が、正しい診断が難しい傾向があるからである。

標的集団にまれな疾患のある被験者が含まれることがあるが、まれな疾患のある被験者を募集することは難しく、費用がかかる。バランスをとれた標的集団よりも、被験者集団にまれな疾患が多すぎるデザインが、適切な場合もあるかもしれない。しかし、そのようなデザインの全体の性能推定値にはバイアスが混入する可能性があり、統計解析プランでこのようなバイアスの可能性を検討すべきである。

8.4 試験の計画、被験者の選択、検体の収集

診断機器は、被験者を直接検査し、その被験者のデータを出したり、被験者から集めた検体の検査を行ったりする。検体を採取してすぐに検査することもあれば、採取してから検査するまで保管する場合もある。既定の治験実施計画書を使用し、その基準を満たした被験者のみから検体やデータを取得する場合、検体や被験者データを前向きに取得すると言う。既定の使用目的がなくアセンブリされた収集物から得た検体や、生体バンクのように別の試験のために定められた治験実施計画書の一部である検体は、前向きに採取されたものと見なされないが、後ろ向き試験に使用できる。同様に、被験者を直接検査する機器（例えばECG、EEG、画像セット）から採取した被験者データは、後で選択、解析するために保存することができる。これも一種の後ろ向き試験である。

前向きに計画された試験では、既定の治験実施計画書を使用する。この治験実施計画書は、使用する選択/除外基準、被験者の募集と選択の方法、検査計画書、解析方法などの試験デザインをあらかじめ規定する。試験期間を通じて、選択/除外基準を満たした被験者が選択される。前向きなプランニングが十分準備されていると、試験集団は標的集団の適切な代表者となり、その試験は使用目的を裏付けるエビデンスを提供することができる。

特定の状況では、前向き試験にバンク検体や過去に収集した被験者データを補足する（例えば、標的疾患が非常にまれで、その疾患の十分な数の患者を前向きに取得することが非常に難しい場合）、又機器の性能を評価するために保管されている検体や被験者データを使用することも認められる。ただし、本指針で述べているバイアスその他の懸念の可能性を適切に解決できることが条件である。バンク検体、過去に集めたデータや画像を試験に追加する場合、検査を行う者や検査結果を解釈する者は、追加した検体やデータを前向きに取得したものと区別できなければならない。

以前保管した検体やデータを後ろ向きに選択する場合は、さらに別の問題が生じる。一部の後ろ向き試験デザインでは、治験責任医師が、有効なデータ、機器が使用する検体、画像その他の保存されている媒体や情報を持っている被験者を探す。後ろ向き選択の例として、三次医療センターへ行って検体を取得する、長期追跡がなされている過去の試験からレジストリデータを使用する、などがある。一般的に、前向きに選択された検体や被験者は、選択プロセスが治験責任医師のコントロール下にある。これに対し、被験者や標本の後ろ向き選択は、例えば、保管検体がある被験者、臨床参照基準の結果がある被験者に限られる。1つの懸念は、後ろ向きに選択された検体や被験者データが標的集団の代表者にならない可能性である（例えば、後ろ向きの検体やデータが標的集団の極端な症例だけを示している可能性がある）。もう1つの懸念は、後ろ向き検体の都合のよい標本は、組織の保管や取り扱い、標的疾患を予測する共変量によって、診断機器の臨床性能をゆがめる可能性がある。前向きプランニングがなく、後ろ向きに取得した検体や被験者データを使用すると、保管されている検体や被験者データを収集した目的（現在の標的集団を代表するものかどうか）、時間経過に伴う検体の劣化のおそれ、被験者データを取得、保管する技術の

変化、保管検体の確実な枯渇など、いくつかの問題が生じる可能性がある。詳細は、残存ヒト検体使用のインフォームド・コンセントに関するFDAの指針を参照のこと。治験依頼者はFDAに相談の上、今ある検体や被験者データが診断機器の使用目的の裏付けに適切かどうかを判断すべきである。

診断臨床性能試験をデザインする際は、検体や被験者データ取得のための治験実施計画書が重要である。IVDの場合、被験者の準備、検体の採取、保管、取り扱いの手順は重要な項目であり、治験実施計画書に詳しく説明すべきである。IVD以外の診断機器は、測定やデータ取得の手順が重要項目である。被験者の測定値や結果をどのように取得するかを、具体的な方法を含め治験実施計画書に説明すべきである（例、特定の刺激処置、電極の留置、データを取得している間の被験者の状態）。

8.5 診断臨床性能比較試験

診断臨床性能試験の目的は、治験機器の性能を実証することである。治験機器と立証済みの機器や方法の診断臨床性能を比較する比較試験は、臨床参照基準を使用する場合のみ可能である。このような試験をデザインする治験依頼者は、デザインの段階で、適切なFDA審査部門と相談することが望ましい。

臨床参照基準がない場合、一致試験で治験機器を別の機器と比較する場合がある。非常に高い水準の一致は、治験機器が既存機器に対して非劣性であると示すことができる。しかし、高い水準の一致は、既存の機器が容認できる性能水準であることが既にわかっている場合に意味がある。

8.6 診断性能試験における盲検化（マスキング）

診断機器の臨床試験には、複数の評価及びユーザー／読み取り者が関与する。例えば、診断性能の臨床試験には、治験機器のユーザー／読み取り者、臨床参照基準の結果を取得する者、場合によっては比較試験で使用する既存機器のユーザー／読み取り者が関与する。治験診断機器のユーザーは臨床参照基準の結果や他の診断評価の結果を知ってはならない（すなわち盲検化されるべきである）。

8.7 機器を使用する人のスキルと行動（全検査コンセプト）

診断機器を使用する際は、専門外の人、瀉血専門医、検査技師、病理学者、放射線医師など、トレーニングやスキルの水準が異なる人が複数の作業を行う。これらの作業には、検体の採取と準備、機器の留置、出力の解釈などがある。トレーニングを受けたスキル、被験者に対する知識、画像及び／又は波形の読み取りの素質を要する仕事の場合、人の能力に差があるのは当然であり、機器の性能に影響を及ぼす可能性がある。したがって、診断機器の臨床性能を評価する際は、治験実施計画書で、機器を使用する人の能力のばらつきについて説明すべきである。特定の機器を使用する人ごとに機器の性能を検証する別の試験が必要なこともある。治験依頼者は試験担当者に行うトレーニングを文書化し、その資料をFDAに提出して審査を受ける必要があるかもしれない。

同じ被験者になされた2つの診断評価の臨床性能比較試験において、比較する評価を、訓練を受けた同じ人が読み取り、解釈する場合、読み取り順のバイアスが混入するおそれがある。このような試験は、様々な出力（例えば画像、スライド）の読み取りを同時にできないので、何らかの既定の順番で行われる。同じ被験者や検体に対して2つの異なる評価が連続した読み取りで行われる場合、1つの評価の認識がもう一つの評価に影響を及ぼすおそれがある。2つ目の評価への影響は、別の評価時に行っても交絡が生じるおそれがある。読み取り順のバイアスを軽減する1つの方法は、

評価と評価の間の時間（「ウォッシュアウト」期間）を長くし、読み取り者の最初の評価の記憶を消すことである。他にも軽減の方法があり、治験依頼者は詳細をFDAの審査部門に相談することが望ましい。

診断臨床性能試験が実施されるという状況が医学的方法を反映していない場合、臨床性能推定値にバイアスが生じる可能性がある。例えば、所定の状況によって標的疾患の罹患率は変動する可能性があり、したがって診断機器の性能の推定値に影響を及ぼす。罹患率が高い場合、読み取り者／解釈者は治験機器の結果を肯定的に見なす回数が多くなり、したがって、診断機器の性能の推定値に影響する。

治験依頼者は、この種のバイアスがその機器の性能の推定値にどのように影響を及ぼすか検討し、できるだけ抑制されるように試みるべきである。

8.8 診断臨床性能試験によくあるバイアスの種類

6.2項で、バイアスを最小限に抑えることの重要性を論じ、前項で診断臨床性能試験に影響を及ぼす可能性があるいくつかのバイアスを紹介した。これらとさらに別の種類のバイアスを以下に詳しく説明する（ただし全てを網羅しているわけではない）。これらを認識し、できるだけ軽減するかなくすようにすべきである。

- **選択バイアス**：試験集団を選択する際の系統的エラー。すなわち試験集団が使用目的集団の代表者ではない。
 - **範囲のバイアス**：選択バイアスの一種で、診断臨床性能の試験が集団のサブグループ間の性能の変動や異質性を相殺できていない。使用目的集団のサブグループ間で性能が異なる場合や、試験が全サブグループを適切に示していない場合に、バイアスが生じる。
- **検証バイアス**：通常は、試験の全被験者に1つの臨床参照基準が使用される。臨床参照基準が試験被験者の1サブセットのみに使用された場合、性能推定値をそれに従って調整しないと、検証バイアスが生じる可能性がある。例えば、検証バイアスは、臨床参照基準の適用の頻度が治験機器の結果によって左右されるときに生じる（例えば、治験機器の結果が陰性の時は臨床参照基準の使用回数が少ないため）。
- **疾患の進行／後退バイアス**：疾患の進行／後退のバイアスは、(a)治験機器と臨床参照基準の結果が同じ被験者で同時に収集されない場合、(b)途中で自発的な回復や疾患のステージの進行があった場合に起こる。現在の健康状態を判断する治験機器の場合、治験機器と臨床参照基準その他試験に使用する診断機器を、1人の被験者にはほぼ同時に使用すべきである。
- **リードタイムバイアス**：診断機器でスクリーニングされた被験者が、リードタイムバイアスによって、診断機器によってベネフィットが得られるように誤って見られる。検査時からの被験者の生存期間は、検査結果がわかっている場合は、結果がわからない場合よりよくなるとは思えないが、早期検出の時間の分だけ後で検出されるより生存期間が増えるため、よくなるように見える。臨床現場によってスクリーニングの期間が異なるときに特に問題となる。
- **時間の長さ選択バイアス（生き残りバイアス）**：標的疾患に長期間かかっている被験者は、短期間の標的疾患がある被験者より臨床試験に組み入れられる可能性が高い。このような患者は通常予後が良好で、標的集団の被験者の代表者にならない可能性がある。その結果、試験被験者の生存期間の推定値が標的集団のものより長くなるおそれがある。

- 外挿バイアス：試験集団の条件や特徴が検査対象のものと異なる。
- 読み取り順バイアス：2つ以上の診断評価を比較するとき、片方の評価の結果の記憶が、読み取り者の解釈に影響を及ぼす。
- 独立した評価がないことによるバイアス：診断機器の性能は、機器の1つ又は複数の側面を開発又は改良するために使用された試験で評価されると、大きくなる可能性がある。診断機器の全ての側面の開発を終えたら、機器の開発に使用された予備試験とは別の新しい試験で、機器の性能を評価すべきである。例えば、1つのデータセットを用いて開発された（「訓練された」）あるアルゴリズムが入った機器の診断臨床性能を、バイアス（「過剰適合」）の可能性を回避するために、別の独立したデータを用いて評価（「試験」）すべきである。

9. 臨床試験の品質の維持

本項では、臨床試験のエビデンスの水準を維持する計画と手法について述べる。これは、臨床転帰試験と診断臨床性能試験の両方に当てはまる。

臨床試験で作成されたエビデンスから、医療機器の安全性と有効性の科学的に妥当な評価が行える。このエビデンスの作成に貢献する主な要因は、試験デザインの選択であり、これによってうまくいけばバイアスの原因も減らすことができる。慎重に試験を行い、データを解析するための健全な科学的方法を使用すると、試験から最大限の情報が得られるはずである。やり方が悪く、分析が不適切な試験は、機器の安全性と有効性を評価するために作成されたエビデンスにあまり頼ることができない。品質試験の実施は、その機器の臨床試験のエビデンス水準を維持するのに有用である。

治験依頼者は、治験が正しくモニタリングされるようにする責任がある（21 CFR 812.40）。臨床試験で作成されるデータや情報の信頼性と有用性を最適化するためには、計画と手法をデザイン作成段階で決定すべきである。この計画と手法は、臨床データの取り扱い、臨床試験の実施、解析戦略のプランニングなどの様々な側面に対処したものであり、試験中に生じるかもしれない変更については、先を見越して説明すべきである。以下にこの側面を詳しく論じる。

9.1 臨床データの取り扱い

Title 21 CFR 812.150(a) に、治験依頼者はある種の情報の完全で正確な報告書を作成し、提出することと定められており、21 CFR 812.140(a) は、各被験者の症例履歴など、ある種の情報の正確、完全かつ最新の記録を要求している。治験依頼者が、臨床試験の開始時に、Good Clinical Data Management Practices（医薬品の臨床試験のデータ管理に関する基準、GCDMP）の原則に従った、データ管理プランと臨床データ取り扱いの包括的なトレーニングプログラムの両方を確立することを、FDA は強く勧めている。多くの臨床試験が電子環境下で実施、報告、記録、解析されつつある今、エビデンスの水準を確立し、バイアスを最小限に抑え、確実に有用なデータを作成するには、データ管理プランが非常に重要である。FDAの規則では、データ管理プランをFDAに提出して審査を仰ぐ必要はないが、FDAは、治験実施計画書にデータ管理プランのエグゼクティブサマリーを入れることを治験依頼者に勧めている。Division of Bioresearch Monitoring のスタッフを含めたPre-Submission会議を要求し、臨床データの取り扱いの最善の方法を話し合うことも、治験依頼者にとって利益となるであろう。

試験データは一貫した形式と構成で収集すべきである。そうすれば、解釈、理解、評価がしやす

くなる。試験、治験実施施設、治験責任医師全体に渡って効率的で標準的なデータ収集の方法を維持すれば、試験全体において高品質のデータを収集できる。さらに、試験に関連したメタデータを比較することによって、試験全体に渡って実施計画の解釈がしやすくなる。データ収集における標準的な語彙と要件の使用は、データ収集が最適化され、データの質と予測性が高まることから、推奨される。

9.2 試験の実施

FDAは、IDEに基づいて行われる臨床試験の進捗報告書を綿密にレビューし、治験責任医師がGCPに基づいた方法で試験を実施しない場合は、それ以上の参加の資格を取り消す権限がある。21 CFR 812.119を参照。GCPや適切なヒト被験者の保護についての指針を含め、臨床試験の実施に関するFDAの規則に対する詳細が利用可能である。

データモニタリング委員会 (DMC) についてのFDAの指針には、臨床試験の依頼者にとってDMCがいつ有用か、またDMCがどのような役割をするかについての情報が記載されている。

これらの臨床試験のプランニングと管理の勧告に従うと、試験のデータの質が向上し、治験によって機器の承認が裏付けられる可能性が増すであろう：

- 無作為化のコードと手順を注意深く保護する。適応無作為化を使用する場合、確率割付の作成に使用するアルゴリズムとデータを保護する
- 試験の盲検性を厳しく維持し、盲検の整合性を評価する。気づいた非盲検化の事象の記録を保持することを、治験依頼者に提案する。
- 治験実施計画書に厳密に従い、マイナーなものも含めてあらゆる種類の治験実施計画書からの逸脱を最小限に抑える。マイナー又はメジャーと見なされる逸脱の種類を治験実施計画書に定義する。治験実施計画書からの逸脱を全て詳細に報告する。大きな治験実施計画書からの逸脱が容認できない割合で起こると、試験結果を作成できなくなる。
- 被験者は、常に、完全に治験実施計画書に従うべきである。被験者のフォローアップ不能の発生と欠損データの影響を減らすために、試験デザインと実施の段階において大きな努力を払うべきである。例えば、治療や有効性の評価を拒否する被験者に安全性のフォローアップを続けるなど、来院なしや連絡漏れの追跡調査の手順を治験実施計画書に明確にする。フォローアップ不能や欠損データの問題に対しては統計手法が使用できるが、このような手法は大幅な想定を使用することが多いので、特殊な試験に対して十分バリデートできない。したがって、フォローアップ不能による欠損データの問題に対する最善の方法は、臨床試験のプランニングと管理の間にその発生を最小限に抑えるプランを作ることである。それでも、欠損データに対しては、感受性解析に加えて適切な統計データ解析法を治験実施計画書であらかじめ決めておくべきである。
- 慎重なデータモニタリングを行い、データが確実で正確であるように徹底し、欠損データを最小限に抑える。臨床試験のモニターのトレーニングや経験に基づいて臨床試験担当者や試験モニターを選択し、ヒト被験者の権利、安全、福祉を含め、試験の質、信頼性、整合性が保証されるようにする。試験モニターは実際の試験（例えば被験者の評価や試験データの記録）に関わってはならない。臨床品質保証プログラムを実施し、試験がデザインと目的に従って行われるようにする。
- 最適な臨床ケア（例；投薬戦略、適切なトレーニングを受けた専門知識のあるオペレーターが機器や管理法を使用する、一貫したフォローアップ手順と戦略）が常に遵守及び／

拘束力のない推奨事項を記載

又は約束されるようにする。

- 統計解析プランにあらかじめ明記されている場合を除き、試験データを注意深く保護し、早期開示によるバイアスを防ぐ。これはオープンラベル試験にも当てはまる。
- 中間解析や治験中止の取り決めが治験実施計画書にあらかじめ定められている、又は安全性の問題に基づいて中止の決定がなされる場合を除き、試験が早期に中止されないような対策を講じる。もちろん、FDAへの申請という目的がなければ試験を早期に中止することはできる。しかし、資金不足のために早期に試験を中止したら、主要解析に失敗したことになる。しかし、治験依頼者は、試験を中止する場合でも、被験者の安全性に関するモニタリングを継続する義務がある。
- 治験依頼者は、機器の試験のトレーニングと経験によって適格性が確認された治験責任医師を選択する。21 CFR 812.43(a)。治験実施施設の担当者（臨床医、試験コーディネータなど）は適切なトレーニングを受けなければならない。治験責任医師のトレーニングを正しく文書化すべきである。

臨床試験デザインと治験実施計画書に、上記の問題点の全てに対処し、最適化し、軽減できる手順を入れるべきである。何らかの臨床試験基準（例えばISO 14155:2011）の遵守を計画している場合、この基準を治験実施計画書に記載する。その場合はこの基準の正しい実施について、治験責任医師と試験モニターにトレーニングを行う。

治験実施計画書からの逸脱に関し、FDAは、参加している治験責任医師が試験デザインの一部の側面に同意していないために、承認された治験実施計画書に従っていないことがわかった。最終的な治験実施計画書が決まる前に、治験実施計画書に関する問題点と逸脱の可能性を解決できるように、治験実施計画書の作成中、治験依頼者があらかじめ治験責任医師と話し合うことをFDAは勧めている。この話し合いによって試験デザインを改善することができる。そうでないと治験実施計画書の逸脱が生じ、試験の解析やデータのプール可能性に問題が生じるおそれがある。さらに治験依頼者は、各参加治験責任医師から署名のある同意書を取得し、それに、その同意書、治験プラン、21 CFR Part 812、審査したIRBやFDAの承認条件に従って治験を実施する治験責任医師の約束を記載しなければならない。21 CFR 812.43(c)(4)(i)。

9.3 試験の解析

やりかたがまずく、不適切な及び／又はポストホック解析は、機器の安全性と有効性の裏付けとなるエビデンスの有用性に悪影響を及ぼす可能性がある。したがって、治験実施計画書に詳細な統計解析プランを既定し、できる限り、試験のデザインでなされた主要な想定を評価するプランを入れるべきである（例えば、クロスオーバーデザインの場合は持ち越し効果、生存解析の場合はハザードの比例性、治験実施施設や地理的領域全体のプーリング解析）。試験終了時のデータ解析では、この試験で作成されたエビデンスの有用性を裏付けるために、この既定のSAPに従うべきである。

予定外のポストホック解析と、治験実施計画書に定めた解析集団からの逸脱は避けるべきである。ポストホック解析の例として、正しい理由なく別の統計解析を使用する、主要評価項目を変更する、解析に既定されていないサブグループを使用する、などがある。このようなポストホック解析は、実験全体の第一種過誤率を増幅させ、それ以外は最適にデザインされ、実施された試験の科学的妥当性を脅かす可能性がある。治験実施計画書には、既知のバイアスの原因に対して推論が堅牢であることを実証するための感受性解析を規定しておくべきである。欠損データが試験の結論に与える影響を批判的に解析することも重要である。

ポストホック解析が明確に説明されており、この種の解析に伴うしかるべき懐疑を持って解釈する限り、この解析が既定の解析を補うことができる場合もある。

9.4 ピボタル試験に予測される変更

中間解析の結果や安全性に関する有害事象の発生により、ピボタル臨床試験の途中で機器の安全性及び／又は有効性を改善するために、機器の変更が必要になる場合がある。例えば、機器のデザインの変更が非常に重大なもので、異なるバージョンの機器で治療を受ける被験者を別の層と見なし、個別に解析して、データを層全体でプールできるかどうかという質問を呈する必要がある。別のインターベンション群の検討を提案する場合は、FDAに相談すべきである。機器の開発段階の後の方で機器のデザインを変更せずにすむように、治験依頼者は、資源のかかるピボタル試験に投資する前に、堅牢な探索段階を利用すべきである。

反対に、被験者集団をプールできるように、試験デザインの途中での変更を計画してもよい。あらかじめいくつかの手直し案を計画し、試験デザインに組み込むことができる。具体的に言うと、試験の統計的整合性を保つために、ベイズ法²⁵ や頻度派中間解析の様々な方法によって、試験デザインの途中での修正（例えば標本数の変更、無作為化の修正）をあらかじめ治験実施計画書に組み込むことができる。試験の結果に応じて試験を修正する順応性のあるデザインをあらかじめ計画することが可能である。この場合、中間結果が出た時に修正を実施するかどうかを誰が決定するか（例えばDMC、独立した統計学者）を定めておくべきである。治験依頼者がベイズ法や順応性のある治験デザインを考えている場合は、できるだけ早くFDAの意見を求めるべきである。手直し案を計画しておかないと、ピボタル試験の科学的妥当性が非常に弱くなるおそれがある。

適切な場合、FDAが予定の被験者コホートのサブセットの段階やその条件で承認し、登録と並行して未解決課題に回答していくこともできる。治験依頼者は、必要な追加情報の入ったIDE補足書類をFDAに提出し、続行が許可されたら登録を拡大できる。その場合、試験デザインは、試験の中断をできる限り少なくして登録を拡大できるように、その前には中間安全性データを見ることができるようになる必要がある。

10. 治験プラン又は治験実施計画書

治験プラン又は治験実施計画書は、臨床試験のデザイン、実施、解析の詳細なプランを記載した文書である（21 CFR 812.25 と21 CFR 860.7(f)(1)を参照）。治験実施計画書には以下の項目を入れる：

- 試験の科学的根拠；
- 評価する被験者集団の定義（選択／除外基準を含む）；
- 提案する機器の使用目的；
- 試験評価項目のリスト；
- 被験者に使用される手順（治療および検査）の記述；
- 使用する適切な統計手法を含めた、試験から導き出されたデータの解析と評価の方法の概要