シリーズ

FDAの一室から

米食品医薬品局医療機器・電磁波製品審査センター
循環器医療機器審査部審査官
内田毅彦

第10回　医療機器の臨床試験

試験の“ダウンサイジング”で鍵を握るバイアスの克服

医療機器と医薬品に違い

医療機器と医薬品ではその開発の特徴が異なるため，臨床試験のデザイニングにも個性のようなものがある。

医療機器の場合，承認を得て日常臨床に使用可能となっても，技術革新に後押しされて次次と新しい機器が生まれてくるため医薬品に比べて圧倒的に短命である。開発者も，絶えず市場に出た後の次世代機器や新規医療機器の開発を考え続けなくてはならない。部品や製造工程の改良といった一部変更も頻繁に行われる。

医薬品でも吸収効率やデリバリー向上のために剤形や投与経路の一部変更は行われているが，医療機器の変更はよりダイナミックと言えよう。医療機器のこうした特徴は，1つの機器開発にかけられる時間と費用が制限されることを意味し，医療機器の臨床試験は，医薬品以上に“ダウンサイジング”が大切なキーワードとなる。

医薬品の治験では，患者の服薬コンプライアンスが試験結果をゆがめる一因になるが，一度の施術で完了する医療機器ではこの心配はない。ただし，医療機器の施術には高い技術を要求するものがあり，ラーニングカーブや医療機関間による治療効果の差を考慮しなくてはならない場合がある。このため，事前に医師トレーニングプログラムを取り入れる治験も見られる。

また，医薬品にプラセボ効果があるように，医療機器の治験にもプラセボ効果があることが知られている。プラセボ効果によるバイアスをなくすため，医薬品のランダム化比較試験（RCT）では二重盲検化が行われるが，医療機器では体内に機器を植え込むといった侵襲を伴うため，二重盲検化が容易でない場合が多い。このため，医療機器の治験では，侵襲を与えない範囲でプラセボ効果を回避するなんらかの工夫を行うことも考えられる。

さらに，コントロール群が既存の治療法である場合，効果の指標に患者や担当医の主観が入る余地があると，新しい医療機器の効果をよいほうに偏って判断する可能性がある（情報・測定バイアス）。被験者が新しい治療法を希望して治験から脱落したり，途中でプロトコル違反を犯して治療法を変更したりする頻度が既存治療群だけに多くなるなどの問題も起こりうる。治験を企画する側も審査する側も，こうした医療機器の治験に特有な問題に注意する必要がある。

代用・複合エンドポイントにメリット

医療機器の治験に限ったことではないが，臨床試験の時間短縮につながる手段の1つに代用（代替）エンドポイント（surrogate endpoint）の利用がある。代用エンドポイントは真のエンドポイント（true endpoint）に代わって用いられる評価項目で，一般に代用エンドポイントの評価期間は真のエンドポイントより短い。

例えば，降圧薬の目的は血圧をコントロールし，動脈硬化の進行を止めることであり，ひいては脳卒中や虚血性心疾患，またはそれによる死亡数減少が真のエンドポイントである[注1)]。しかしこの評価には長い年月がかかるため，血圧低下という代用エンドポイントを使うことで試験期間の短縮を図ることができるようになる。

ここで重要なのは，代用エンドポイントは真のエンドポイントを正しく反映していなければならないことである。降圧薬の例で言えば，血圧の低下によって脳卒中や虚血性心疾患の発生が減少することが，適切な臨床研究に基づいたエビデンスとして確立していなければならない。仮に，ある薬剤がプラセボに比べて統計学的に有意に血圧を下げたとしても，降圧の結果が真のエンドポイントに影響しない程度であれば，妥当な代用エンドポイントと言えないのである。ちなみに，有害事象の発生をエンドポイントとするRCTにおいては，代用エンドポイントを用いたほうが被験者数も少なくできる場合もある[注2)]。

また，複合エンドポイント（composite endpoint）を利用することでも被験者数を縮小できる。例えば，心不全治療における究極のエンドポイントは心臓病関連死であるが，これだと軽症心不全の場合では発生するイベント数が多くならない。しかし，心不全患者は状態がよくないと発作を起こし入退院を繰り返すため，入院回数や次回入院までの日数をエンドポイントに加えると総イベント数が増え，統計学的検出力が高まるので，必要な被験者数が軽減される。

複合エンドポイントのもう1つのメリットは，1つのエンドポイントだけでは見落としてしまう結果（競合リスク）を回避できることである。先に挙げた例で，入院回数のみを代用エンドポイントとして利用する場合，死亡例ではもはや入退院数は増えないため，死亡という最悪の結果が多く含まれた群が逆によくなってしまう可能性が出てくる。しかし，ここでエンドポイントに「死亡」を含んでいれば，そのリスクは回避できる。ただ，臨床的に“ハード”な「死亡」と“ソフト”な「入退院」を同じエンドポイントで計ってしまう問題もあるため，これらの結果は必ず報告書へ個別に記載される必要がある。

同じ例で，ハードエンドポイントとソフトエンドポイントの臨床的重みの違いを重視して，心臓死と次回入院の2つのイベントを別々のエンドポイントとし，そのいずれかで有意差があればよしとするデザインを考えてみよう。この場合は，第一種の過誤（Type 1 Error，本来真である仮説をたまたま得た結果によって棄却してしまう誤り）がきちんとコントロールされているかが問題になる。①心臓死②次回入院それぞれに仮説を立てる場合，それぞれ有意水準を0.05とし，どちらか一方の試験でP値が0.05未満なら有意差ありと判断するには落とし穴がある。P値0.05は5％の確率でその結果が起きたという意味だが，2つの仮説がこの値を満たす確率は，1－（0.95×0.95）＝0.0975となり，どちらかがP値0.05未満ならよしとすることは9.75％の偶然を容認してしまうことになる。ここで，第一種の過誤をコントロールするには，①心臓死で有意，かつ②次回入院で有意とするか，それぞれの有意水準を0.025にするといった方法が必要になる。

この第一種の過誤のコントロールの問題は，もともとプロトコルに含めていなかったエンドポイントを試験終了後に追加し，それを新しい治療法の有用性として報告する誤りにも通じている。心不全の例で言うと，心臓死においても次回入院においても有意差が認められなかった際に，データを調べ直した結果，不整脈の発生抑制で有意差を見つけて申請者が不整脈の適応を要求しても，それがあらかじめプロトコルに検証する仮説として記載されていなければ，第一種の過誤の問題からFDAは原則としてこれを認めない。

ヒストリカルコントロールとOPCを共有

ヒストリカルコントロール（対照群として過去の試験を利用）を用いた試験も，被験者を少なくできる手法である。また，既にコントロールとして確立されたデータがある場合，それをOPC（objective performance criterion：複数の臨床試験結果のメタ解析による結果をクリアすべき基準とする）として共有できれば，複数の企業が開発に利用できるうえ，医療機器間の整合性が取れるという利点もある。

しかし，絶えず新しい治療法が模索される医療機器の世界では，そのOPC自体がすぐに時代遅れになる可能性がある。また，OPCをつくる際，もとになる臨床試験間で対象集団が均一であるか否かが問題になるし，新しい治療法の被験者群がOPCの対象集団に一致するかも問題になる。エンドポイントに影響を与える外的因子のアンバランスが結果をゆがめることになるからである。

こうした点から，ヒストリカルコントロール試験はRCTに劣り，ランダム化するには倫理的・実務的に問題があるような場合に，バイアス軽減への十分な配慮をしたうえで用いられることが多い。ただ近年は，Propensityスコア解析[注3)]というバイアスの軽減につながる方法も用いられ，ヒストリカルコントロールを用いる試験の有用性を高めている。

メリーランド州ロックビルにあるFDA本部。現在，同州ホワイトオークにある政府所有地への移転作業が進んでおり，2011年の移転完了時には各研究所を含むFDAのおもな機能が1か所に集約される

注1）虚血性心疾患や脳血管障害を発症しても軽症で寿命や日常生活に影響を与えない場合，そのイベントをエンドポイントとして死亡例と同じに扱うと重みに差が出るという見方もできる。一方で軽症循環器合併症はその後により重篤な事象が起こるリスクが高いことからエンドポイントに含めてもよいという見方もある。エンドポイントは医学的見地から選択すべきであることはもちろんだが，それが試験のデザインを大きく左右するため，実際的，統計学的見地を踏まえて総合的に決める必要がある。

注2）同じ相対危険度ならば，両群の総イベント数を総被験者数で割った数が0.5に近いほど統計学的検出力が高いため。例えば，冠動脈ステントの臨床試験で真のエンドポイントを循環器関連死・心筋梗塞発症とし，5年で治療群・対照群のイベント割合がそれぞれ0.03，0.05〔オッズ比（OR）1.7，n＝500/群〕だったとすると，α＝0.05でパワー30％。これに対し代用エンドポイントをステント再狭窄率とし，1年で治療群・対照群のイベント割合がそれぞれ0.15，0.23（OR 1.7，n＝500/群）とするとパワーは90％にもなる。

注3）ヒストリカルコントロール群と治療群の間に，結果に影響を及ぼすほかの因子があるとその影響を受けるため，ある患者がどれくらい他の因子を含んでいるかをスコア化し，そのスコアが近い患者同士を選んで比較しようとする考え。例えば冠動脈ステントであれば，糖尿病の合併はスコア3点，過去のステント治療歴があると1点，病変長は長いほど悪影響なのでA mmだとスコア2×A点というスコア化をあらかじめ多重回帰分析などの方法で行い，スコアに応じてマッチングする。