ライティング徒然草

エグゼクティブ・アドバイザー　林　健一

第 10 回 「有用性の総合評価」って何だ?

2019 年にライティング徒然草を書き始めてから早くも 2 年が経過し，今回が連載 10 回目となる。ちょうど区切りがよいので，ここで国内の治験の評価方法や解析方法がどのように変化したのかを振り返ってみたい。これらは 1990 年代の半ば以降に大きく変化したが，製薬企業内で世代交代が進んだため，当時を知る開発担当者はほとんど残っていないのが現状である。しかし，かつての問題点を理解することは，今後医薬品を開発するうえで何かの役に立つかもしれない。こうした背景から，記憶を呼び起こそうと思い立った次第である。

今回は全般改善度・概括安全度・有用度という総合評価の尺度を紹介する。これらは国内の治験のみで用いられ，ICH E9 ガイドライン「臨床試験のための統計的原則」の 2.2.4 項「総合評価変数」でその問題点が指摘されている。しかし，現在ではこれらを使用することがないため，多くの方はこの項を読んでも意味を理解できないであろう。たしかに，こうした尺度を用いることは今後もないであろうが，ここには総合評価の普遍的な問題点が示されている。このため，各尺度の内容と，これらが用いられた背景を以下に記すこととする。

まず，全般改善度というのは，個々の患者の自覚症状や検査値の変化を総合して，その患者に対する最終的な治験薬の有効性を判定したものである。具体的には，「著明改善・中等度改善・軽度改善・不変・悪化」という 5 段階で患者ごとの改善度を判定した後，中等度改善以上と判定された患者の割合を群別に記述するのが一般的であった。なかには，「悪化」を「軽度悪化・中等度悪化・著明悪化」に細分化し，合計 7 段階で判定する試験も存在したが，この場合でも中等度改善以上と判定された患者の割合を群別に記述したので，本質的な違いはない。

次に，概括安全度は，有害事象の有無や臨床検査値の変化を総合して，個々の患者に対する最終的な治験薬の安全性を判定したもので，ほぼ全試験で「(安全性に)問題なし・やや問題あり・問題あり・極めて問題あり」という 4 段階のカテゴリーが用いられた。この尺度では，有害事象や臨床検査値の異常変動が認められなければ「安全性に問題なし」と判定し，軽度・中等度・高度の有害事象が発現した場合には，それぞれ「やや問題あり」「問題あり」「極めて問題あり」と判定するのが一般的であった。

最後の有用度は全般改善度と概括安全度を統合したもので，「極めて有用・有用・やや有用・有用性なし・好ましくない」という 5 段階を設けるのが一般的であった(「好ましくない」を「やや好ましくない・好ましくない・極めて好ましくない」と細分化する試験も存在した)。評価の対象となる患者の概括安全度が「安全性に問題なし」と判定された場合，有用度の判定結

果はおおむね全般改善度の判定結果と一致した。すなわち，全般改善度が「著明改善」と判定されれば，ほとんどの場合，有用度も「極めて有用」と判定された。一方，概括安全度が「やや問題あり・問題あり・極めて問題あり」と判定された場合には，判定結果に応じて，有用度が全般改善度よりも 1～3 段階低く判定されることになった。そして，有用度がその患者に対する最終の判定，すなわち主要評価項目とされたのである。

これらの評価尺度は担当医師が自身の主観に基づいて判定した。自覚症状や定量データの変化を一定のアルゴリズムに従って総合することはなく，すべてが「担当の先生のご判断」という名前のブラックボックスの中で判定されたのである。この点で，全般改善度等は，米国リウマチ学会 American College of Rheumatology が設定した関節リウマチの評価尺度 ACR20 などとは決定的に異なる。このため，判定結果の妥当性や信頼性を担保することができず，日本で実施された治験の結果を欧米の規制当局が受け入れるうえで大きな障壁となった。なかでも，槍玉に挙がったのが有用度である。

有用度は，いわばベネフィットとリスクを統合した尺度である。しかし，ベネフィットとリスクを統合する必要があるとしても，それは想定するターゲット集団を対象として考えるべきものであり，個々の患者内で統合する必要はない。たとえば，ある患者の症状や徴候が劇的に改善した結果，全般改善度が「著明改善」と判定されても，概括安全度が「安全性に問題あり」と判定されれば，有用度は「やや有用」と判定されることになる。一方，症状や徴候の改善が軽度で，全般改善度は「軽度改善」と判定されたが，安全性には問題がなかった場合にも，有用度は「やや有用」と判定される。すなわち，最終判定が同じになってしまうが，これらの患者に生じたことは大きく異なる。この結果，E9 は 2.2.4 項で「有用性の総合評価」に言及し，有用度に引導を渡したのである。

では，なぜ全般改善度等の尺度が用いられるようになったのであろうか。実は，これらを提案したのは，国内でランダム化二重盲検比較試験を実施することの必要性を訴えていたグループである。このグループの中心となったのは一人の精神科医で，抗うつ薬の臨床評価を専門としていた。抗うつ薬の有効性は単一の症状の変化から評価することができず，どうしても複数の症状の変化を総合して評価することが必要となる。こうした背景から，全般改善度等を思いついたものと推定され，このグループの影響が広まるにつれて総合評価の尺度が定着していった。

さらに，総合評価の尺度を使用せざるを得ない事情もあった。それは，コンピュータの演算処理速度がもたらす制約である。当時は，臨床試験のデータを解析しようとすると，どうしても大型計算機（当時はメインフレームと呼ばれた）に接続する必要があった。そうしないと，解析に膨大な時間がかかったのである。しかし，大型計算機はいつでも自由に使用できるわけではない。たとえば，製薬企業の解析担当者が自社の大型計算機を使用する場合には，経理処理等の業務が空いた時間を狙って解析を実施する以外に方法がなかった。当時の製薬企業にとって，臨床試験の統計解析は最優先の業務ではなかったのである。

こうした制約下では，どうしてもバッチ処理が必要となる。すなわち，実行するプログラムを事前にすべて確定したうえで，大型計算機にデータを流し，一気にプログラムを実行するのである。この処理をするためには，対象疾患や薬剤の特性に関係なくデータ構造を統一するとともに，プログラムも統一するほうが都合がよい。具体的には，データ全体を①患者背景に関する定性・定量データ，②全般改善度等の総合評価，③反復測定する定性データ（例：自覚症状の重症度），④反復測定する定量データ（例：血圧の測定値や臨床検査値）に分け，各データに用いるプログラム（記述統計量の算出，仮説検定の実施）を統一すれば，統計に詳しくないものでも処理が可能となる。実際，二重盲検試験の必要性を提唱するグループは解析にバッチ処理を用いていた。処理をしたのは，このグループの事務局である。

以上が総合評価の尺度が生まれた背景と，これらを必要とした事情である。現在の価値基準から考えると，統一された評価尺度を用いたことは奇異に見えるかもしれない。しかし，当時は「経験を積んだ医師が評価すれば，薬が効くか効かないかは一発でわかる。対照群を設置した二重盲検試験などを実施する必要はない」という意見が医師の間で大手を振ってまかり通っていた時代である。疾患領域にかかわらず，二重盲検試験の必要性に理解を示す医師は少数であった。こうした状況下では，評価尺度を統一することにも一定の意味があった。統一によって，同じ様式で二重盲検試験を実施することが可能になり，先に二重盲検試験を経験した医師が領域をまたいで実施方法を指導するようになったのである。

コンピュータの演算処理速度が飛躍的に向上し，SAS を用いて対象疾患や薬剤の特性に応じた解析プログラムを書くことが可能となった現在では，全般改善度等の評価尺度を用いる必要性はどこにも存在しない。しかし，だからといって「昔の臨床試験のやり方はいいかげんだった」と一刀両断に切り捨てるのも適切ではない。実は，ここには貴重な教訓が隠れているのである。それは，評価尺度の妥当性や信頼性を検証するということである。

現在でも，総合評価の尺度を使用しなければならないことは多い。たとえば，統合失調症やうつ病といった精神疾患に対する治験薬の有効性を評価する場合には，多様な症状をスコア化して総合せざるを得ない。健康に関連する Quality of Life（QOL）を評価する際も同様で，様々な質問に対する患者の回答を総合しなければ QOL を評価することはできない。ただし，これらの評価尺度を用いる際には，質的な研究を実施した後，定量的な手法を用いて評価方法の妥当性や信頼性を検証することになっており，この検証手順はほぼ確立しているといってよい。

全般改善度や有用度の問題点はこうした手順を踏まなかったことで，この点を理解しないと，将来同じ過ちを繰り返さないとも限らない。約 20 年前，E9 を検討する過程で生じた日本に対する海外の集中攻撃を二度と繰り返してはならないのである。

追記：

先のグループの中心となった精神科医は，当時の精神科領域を代表する 2 名の医師を前

にしてこう言ったそうである。

「経験を積んだ医師は薬の有効性を評価できるとおっしゃるのでしたら，先生方が一番効くと考える抗うつ薬を今ここで挙げてみてください」

こうして，２名の医師が同時に発言するように仕向けた後，それぞれが異なる抗うつ薬を挙げるのを確認し，二重盲検試験の必要性を説いたのである。

「ほらね。やっぱり二重盲検試験は必要でしょ」。

こうして，精神科領域でも二重盲検試験を実施できる基盤が整った（注：ご本人から直接伺った話である）。ただし，プラセボを対照薬とすることにはなかなか理解が得られず，その結果，海外で広く使用されている抗うつ薬が国内では使えないという悲劇が長く続くこととなった。

けれども，これは別の物語。いつかまた，別のときに話すことにしよう。