第8回 データサイエンスラウンドテーブル会議

テーマ：非統計家に対する統計的な情報の提示について

2024年6月25日(火) @PMDA会議室

非統計家に対する統計的な情報の提示

- 臨床研究の研究デザインや解析方法に基づく医薬品についての正確な情報を伝達は，生物統計家の役割の一つ
- 非統計家の背景知識を考慮しながら，誤解を生むことなく正確な情報を伝達することが重要
  - 例えば，“医療用医薬品製品情報概要等に関する作成要領” には掲載可能なp値の分類が明記されている，その統計的背景の理解なしには必要な情報を正確に伝えることは難しいであろう
- 本日のテーマ：
  - 昨今試験デザインや解析手法の複雑化が進む中で，医師向けに臨床研究結果を説明する状況を想定し，結果の解釈に必要な統計的背景に関し統計的な情報を提示する際にどのような課題や限界があるのか，そしてどのような対処が考えられるかについて，2つの具体例とともに議論・整理する

議題

1. 各事例の研究結果を解釈するうえで必要だと考えられる（非統計家に対して確認すべき）統計的な背景知識を議論・整理

2. 事前配布資料を基に，非統計家への正確な各事例の研究結果の情報伝達に必要な情報の過不足について議論・整理
   - 一つの目の議題で整理した，必要な背景知識の有無別での議論・整理

3. ほとんど統計知識のない方（患者さんなど）向けの情報提供を想定し，誤解のない提供方法を各事例をもとに議論・整理

事例1：
検証的試験における評価項目別のp値提示方法

アトピー性皮膚炎患者を対象としたバリシチニブの有効性及び安全性を評価する第III相試験（BREEZE-AD1）

試験概要

目的	中等症から重症のアトピー性皮膚炎患者を対象に、オルミエント1mg、2mg及び4mg1日1回投与の有効性をプラセボと比較検証するとともに安全性を評価する。
対象	外用治療で効果不十分又は不耐である中等症から重症のアトピー性皮膚炎患者624例
試験デザイン	多施設共同、無作為化、二重盲検、プラセボ対照、並行群間比較検証試験
解析計画	両側有意水準5%で全体的なfamily-wiseの第1種の過誤確率を制御するために、主要評価項目及び主な副次評価項目について、グラフィカルアプローチを用いて多重性を調整した解析を行った。主要評価項目はco-primary評価項目であり、ある用量において、IGAスコア≦1達成率及びEASI-75達成率がともに統計学的に有意である場合に、その用量における優越性が検証されたとした。ロジスティック回帰分析を用いて、離散変数に関する治療比較を行った。

議題1で出た意見

研究結果を解釈するうえで必要だと考えられる統計的な背景知識

- 多重性
  - Family-Wise Error Rate (FWER)
- Co-primary
  - 主要評価項目は1つであることが多い
- グラフィカルアプローチ
- P値
  - 定義・検証と名目上のp値の区別
- 統計モデル
  - 統計モデルを使用している理由（共変量の調整のためなど）

議題2で出た意見

非統計家への正確な各事例の研究結果の情報伝達に必要な情報

- 多重性
  - これを解決するためにグラフィカルアプローチを行う、というぐらいの説明をする
  - 多重性についての詳細は聞かれたら説明する
- グラフィカルアプローチ
  - 基本的には、結果の理解ができればよい
    - 詳細な説明はしないが、多重性の調整を行ったことや大まかな概念は説明する
    - グラフを出すと、ここまで検証しました、ということが言いやすい
      （上から順番でここまで！と優先順位が伝えやすい）
  - 必要に応じて計画段階（グラフィカルアプローチのモチベーション）を説明する
    - 重要な仮説，研究全体での制御したいFWER，試験成功の基準を明確にする
  - エンドポイントの順番の根拠（臨床的観点から）も必要に応じて説明

議題2で出た意見

非統計家への正確な各事例の研究結果の情報伝達に必要な情報

- P値
  - P値を出すのではなく、有意水準との大小のみを出すのも一案
    - 小さいp値の時によく効くんだと誤解があることからは大小だけ出すことはよい
    - 有意水準との比較のみを出すと、推定値を気にするきっかけになるかもしれない
    - 一方で低いP値を出す方がインパクトはあるかもしれない
  - 推定値が小さい時にはP値が小さい方が伝えやすいのか
  - 二つの差の違いを区別　推定値の大きさ⇒臨床的な差の大きさ
p値の小ささ⇒統計的な差の大きさ
  - PrimaryがP値で判断できるのは、検出力まで考慮しているからだが、副次では未考慮
    - Over powerの可能性あり
    - 被験者数設計時に考慮した評価項目以外の検出力の記載はできればしたくない。。。
    - 検出したい効果量であったか議論すれば、結果的に検出力の議論と同様

議題2で出た意見

非統計家への正確な各事例の研究結果の情報伝達に必要な情報

- P値
  - 多重性調整のない仮説の結果まで提示する必要があるのか
    - ルールがないといい結果のみ出すことになるのではないか
    - 全部見たい、網羅的に載せておくという視点もあるか
    - 基本的には出さない方がよい。調整している結果への信頼感がなくなる
  - 「検証された」と「有意に高かった」の区別でどれくらいわかるのか
    - 統計家だけが区別を理解していても，読み手は理解していない可能性がある
- 統計モデル
  - 0.025に分けたこと（αの調整）を説明するが、その先の詳細や、背景、モデルの仮定は話さない

議題3で出た意見

統計知識のない方向けの誤解のない情報提供方法

- 統計知識がない方が知りたいこと
  - 薬が効くのか、効かないのか
  - 効果の大きさ
    - Primaryに限らず、症状改善やQOL改善の割合について伝える
      - 1か月飲んだら〇%の人が治る！
      - 〇%治療効果がある！
  - 安全性
  - 集団での効果ではなく自分自身での効果
- 効果があることの根拠までは詳しく知らなくてもよいが、情報の取捨選択は患者本人がするので、それなりに情報提供したほうがよい

事例2：
早期中止判断のための中間解析を伴う試験における結果の提示方法

早期AD患者を対象としたアデュカヌマブの有効性及び安全性を評価するための2つの第III相試験（EMERGE試験及びENGAGE試験）

EMERGE試験及びENGAGE試験の概要

- 二つの同一デザインの第III相試験

項目	説明
目的	早期AD患者（ADによる軽度認知障害（MCI）及び軽度AD型認知症）におけるアデュカヌマブの有効性及び安全性を評価
対象	ADによるMCI又は軽度AD型認知症の臨床基準を満たし、アミロイド病変が確認された50～85歳の患者
試験デザイン	ランダム化、二重盲検、プラセボ対照試験
投与群	低用量群、高用量群、又はプラセボ群に1:1:1の比でランダム化
主要評価項目	CDR-sum of boxes (CDR-SB)（スクリーニング、26、50、78週で評価）
投与期間	76週
用法・用量	次のスライドに記載
中間解析の結果	いずれの試験も中間データの無益性解析の結果に基づき早期に中止された

最終解析の結果

- EMERGE試験では主要評価項目が達成されたが、ENGAGE試験では達成されなかった

Table 2. Primary and secondary endpoints at week 78

Endpoint	EMERGE			ENGAGE		
	Placebo decline ± SE (n=548)	Difference vs placebo (%) | | 95% CI P		Placebo decline ± SE (n=545)	Difference vs placebo (%) | | 95% CI P	
		Low dose (n=543)	High dose (n=547)		Low dose (n=547)	High dose (n=555)
Primary						
CDR-SB*	1.74±0.11	−0.26 (−15%)	−0.39 (−22%)	1.56±0.11	−0.18 (−12%)	0.03 (2%)
		−0.57, 0.04	−0.69, −0.09		−0.47, 0.11	−0.26, 0.33
		.090	.012		.225	.833

出典：Haeberlein SB et al., 2022 Table 2参照

統計的な知識があまりないドクターへの説明を想定

議題1で出た意見

研究結果を解釈するうえで必要だと考えられる統計的な背景知識

- 中間解析についての知識
  - 有効中止（多重性の調整）と無益性中止（条件付き検出力）
  - 多重性の調整に関する知識
- P値に関する知識
  - 製品情報概要の「名目上のp値」

統計的な知識があまりないドクターへの説明を想定

議題2で出た意見

非統計家への正確な各事例の研究結果の情報伝達に必要な情報

- AD対象の2試験(EMERGE, ENGAGE)
  - 前相の臨床試験情報、疾患情報
    - FDA以外、各国の承認状況
  - 2つの試験が存在
    - 両試験の結果に齟齬があったこと（無益性中止後に1試験のみ最終解析met）、条件付き承認になったことは強調すべき
      - なお条件付き承認の理由はバイオマーカーへの効果が認められたこと

統計的な知識があまりないドクターへの説明を想定

議題2で出た意見

非統計家への正確な各事例の研究結果の情報伝達に必要な情報

- 治験薬によるBenefitの臨床的意義（スコア0.5の差）
- 安全性の情報（Risk/Benefitを考えるため）
- エンドポイントとなっているスコアに関する特性（CDR-SB）
- 脱落症例とその理由
- 差だけではなく各群のスコアも確認する

統計的な知識があまりないドクターへの説明を想定

議題3で出た意見

統計知識のない方向けの誤解のない情報提供方法

- **条件付き検出力**
  - なぜ中止したのか、を理解してもらう
  - そもそも「検出力」とは
    →試験の成功確率、検定で有意になる確率
    治療効果があると仮定し、試験を100回実施したときに、80回で試験成功となる（Power 80%）
  - 中間解析時の結果を踏まえた上で、試験の成功確率を解析したものが「条件付き検出力」
    「このまま試験を継続した際に、薬剤の有効性が示される確率」
    20%の基準に関しては、慣例的な設定に基づいている
  - 2試験の結果をPOOLした上で条件付確率を算出している
    （2試験の有効性が同じであるという仮定に基づく）

- **試験の解釈（そもそもこの試験は成功か、失敗か）**
  - 統計的な疑義はある（試験成功 or 失敗の二値判断はできない）が、FDAは「承認」という結論を出しているので、その判断について説明しなくてはいけない
  - 難しい疾患領域で、他剤の状況なども踏まえて承認されたと予想される
  - FDAでも議論になった判断であり、自由診療などの背景も踏まえての判断であった可能性がある

統計的な知識があまりないドクターへの説明を想定

議題3で出た意見

統計知識のない方向けの誤解のない情報提供方法

- **中間解析**
  - 有効中止のための検定と無効中止の基準について説明する
  - 中間解析時と最終解析時において、異なる集団であると予想されるため（登録基準の変更）、中間解析で中止が判断された状況でも最終解析を実施することにしている
- **多重性**
  - αエラー5%の意味（差がない時に差があると判断してしまう確率が5%）
  - 複数回検定すると、どこか一回でも偶然差があると判断される確率が5%を超えてしまうため、調整する必要がある
  - P-valueを単純な閾値として認識しているドクターが多いため、そこから説明する必要がある
- **無益性中止の解釈**
  - 「統計的に薬剤の有効性が無い」ではなく、「このまま試験を継続しても、薬剤の有効性が示される見込みが低い」という解釈
  - プロトコル改定のため「このまま試験を継続した」わけではないため、無益性中止の判断時と異なる集団で最終解析が実施され、有意差がついた
  - プロトコル改訂により集団特性を変更することは統計的に問題があるため、試験結果の解釈を難しくしている

統計的な知識があまりないドクターへの説明を想定

議題3で出た意見

統計知識のない方向けの誤解のない情報提供方法

- そもそも2試験やっている理由・解釈を説明する必要がある
  - FDAの法令
  - 通常は2試験ともメットした際に承認される