

当日のスケジュール

時間	内容
10:30—11:15	講義：単群試験の薬効評価法と統計的課題の紹介（兼清）
	講義：新薬開発における単群試験の審査経験（浅野）
	講義：レジストリデータの現状と新薬開発における課題（嘉田）
11:15—12:30	ディスカッション1：仮想ケースを参考にした単群試験による薬効評価において生じる統計的課題の対応策の議論
	ディスカッション1：各グループからの発表
12:30—13:30	昼休憩
13:30—16:00	ディスカッション2：単群試験による薬効評価が許容される状況について
	ディスカッション2：各グループからの発表
	ディスカッション2：パネルディスカッション（兼清〈産〉、嘉田〈学〉、浅野〈官〉）

当日の議論の状況に応じて、当初のスケジュールとは変更になっている。

ディスカッション1

ディスカッション1の目的

- ディスカッション1では仮想ケースを基に、単群試験による薬効評価をどのように実施すべきかについて議論を行った

仮想ケース(1/4)

- 疾患対象:希少疾病X
 - 小児期に発症し、年齢と共に状態が悪化していく
 - 開発薬剤は状態の悪化を止めることが目的
 - 症例数は実施可能性の観点から国内で20例程度が最大
 - 既承認の薬剤は国内外で共にない
 - 開発薬剤は国内のみで開発され、国際共同治験は困難
 - プラセボ対照の比較試験は困難な状況
- 有効性の評価項目
 - 被験者の状態及び医師による全般評価等を数値化し、その数値の合計値のベースラインからの変化量(24週後)
 - 合計値が上昇するほど悪化
 - どの程度の変化があれば臨床的に意義があるかは不明
- 無治療時の有効性の評価項目の値について、症例報告及び国内レジストリーデータが存在

仮想ケース(2/4)

症例報告

- 論文A:
 - 海外患者のみのデータ
 - データの年齢の平均は低め
- 論文B
 - 国内患者及び海外患者のデータ
 - 低年齢から高年齢まで。
- 論文C
 - 国内患者のみのデータ
 - データの年齢の平均は高め
- 有効性の評価項目の要約統計量(平均±標準偏差)
 - 論文A: 20 ± 20 (N=10)
 - 論文B: 15 ± 15 (N=15)
 - 論文C: 10 ± 30 (N=5)

仮想ケース(3/4)

国内レジストリーデータ

ID	Xの変化量	年齢	性別	体重
1	20.0	0.3	F	4.0
2	10.0	1.0	M	8.0
3	30.0	3.0	F	15.0
4	7.0	13.0	M	35.0
5	10.0	3.0	F	-
6	30.0	0.5	M	3.0
7	20.0	1.0	F	8.0
8	-	11	M	30.0
9	30.0	0.6	F	4.0
10	-	1.6	M	9.0

ID	Xの変化量	年齢	性別	体重
11	50.0	3.0	F	17.0
12	20.0	3.0	M	-
13	10.0	0.5	F	-
14	8.0	9.0	M	25.0
15	-	12.0	F	30.0
16	20.0	4.0	M	-
17	10.0	0.5	F	3.0
18	30.0	0.5	M	4.0
19	20.0	2.0	F	14.0
20	10.0	1.0	M	9.0

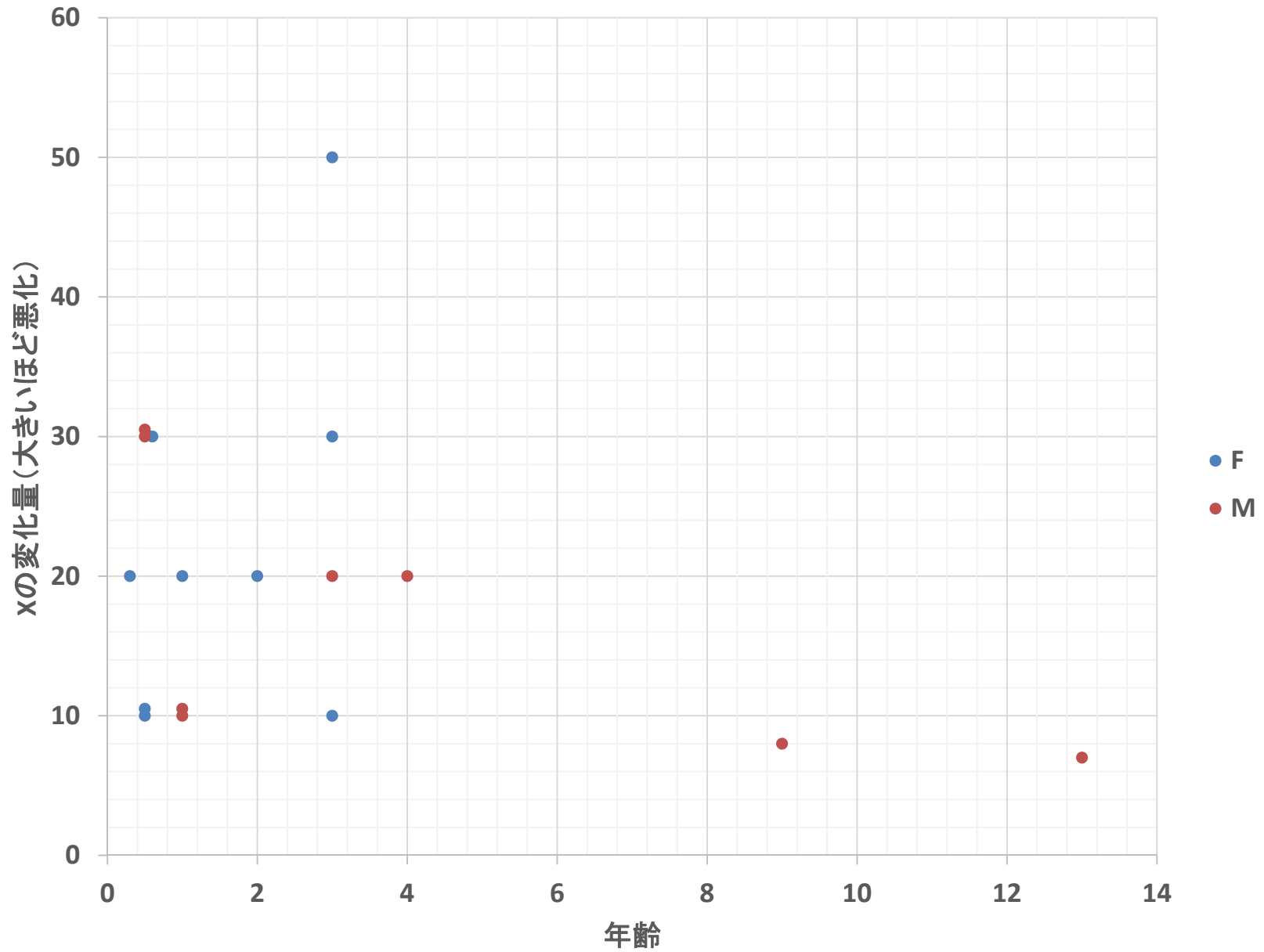
仮想ケース(4/4)

論点

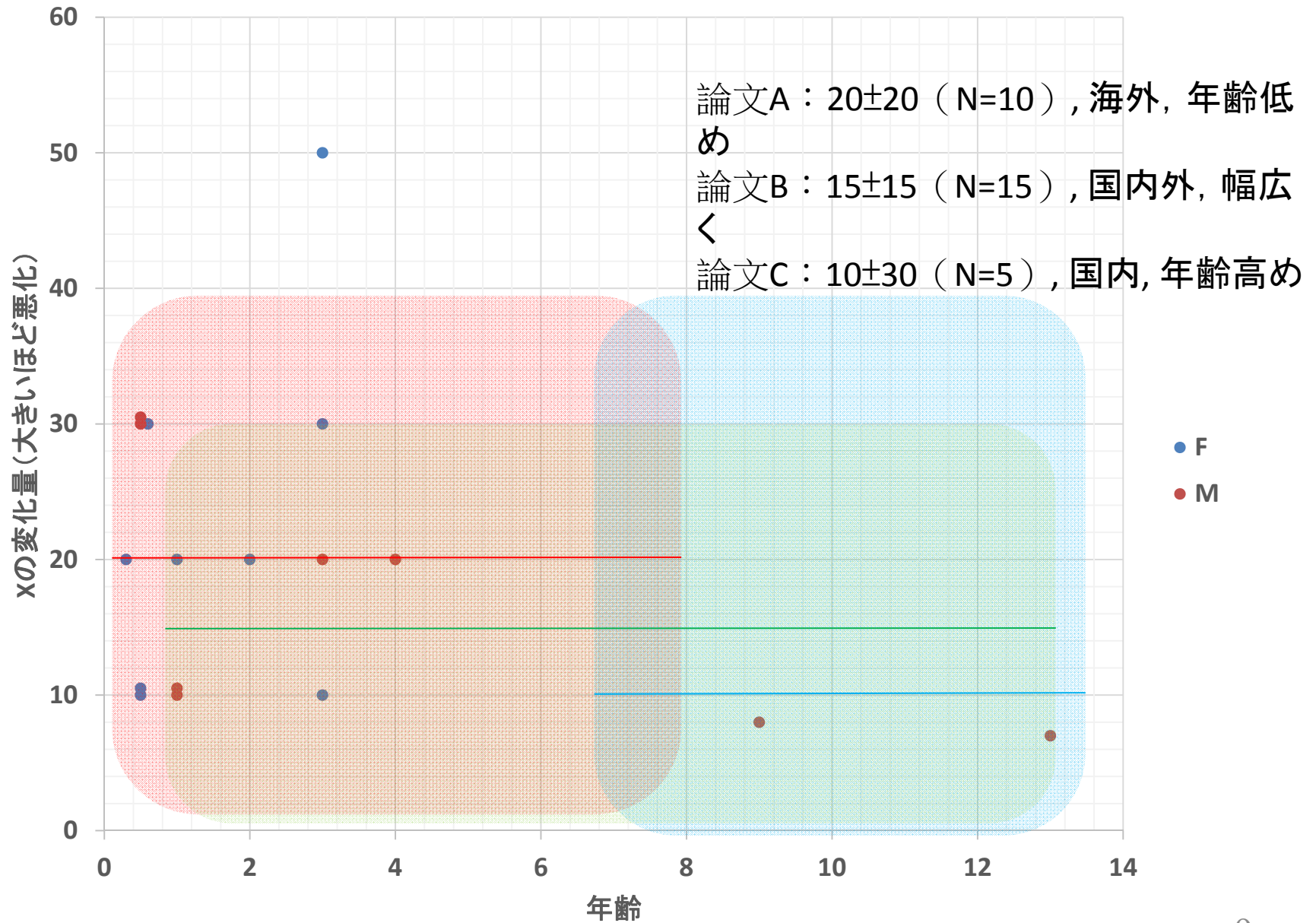
- 閾値を設定して有効性を評価する場合
 - 閾値の設定方法
 - 有効性の解析方法
 - 試験デザイン又は解析において、バイアスを最小化する方法
 - 当該評価の良い点・悪い点
- レジストリデータを外部対照群として有効性を評価する場合
 - 有効性の解析方法
 - 試験デザイン又は解析において、バイアスを最小化する方法
 - 当該評価の良い点・悪い点
- 上記以外のフレームワークの利用可能性
- 各テーブルで最適だと思ふ方法を1つ選択し、その選択理由及び方法の詳細をまとめてください。

仮想ケースの参考資料(1/5)

年齢とxの変化量

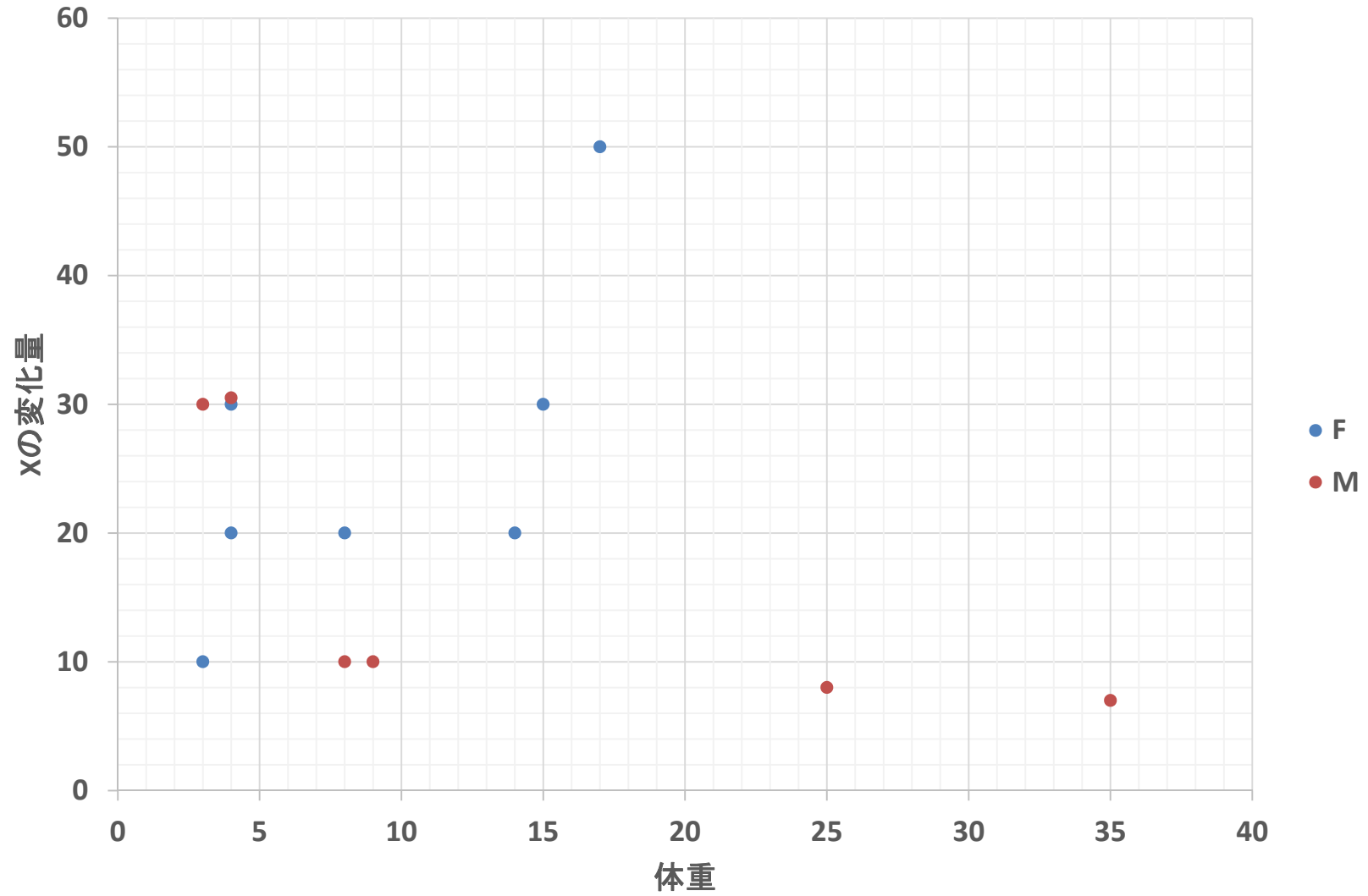


仮想ケースの参考資料(2/5) 年齢とxの変化量



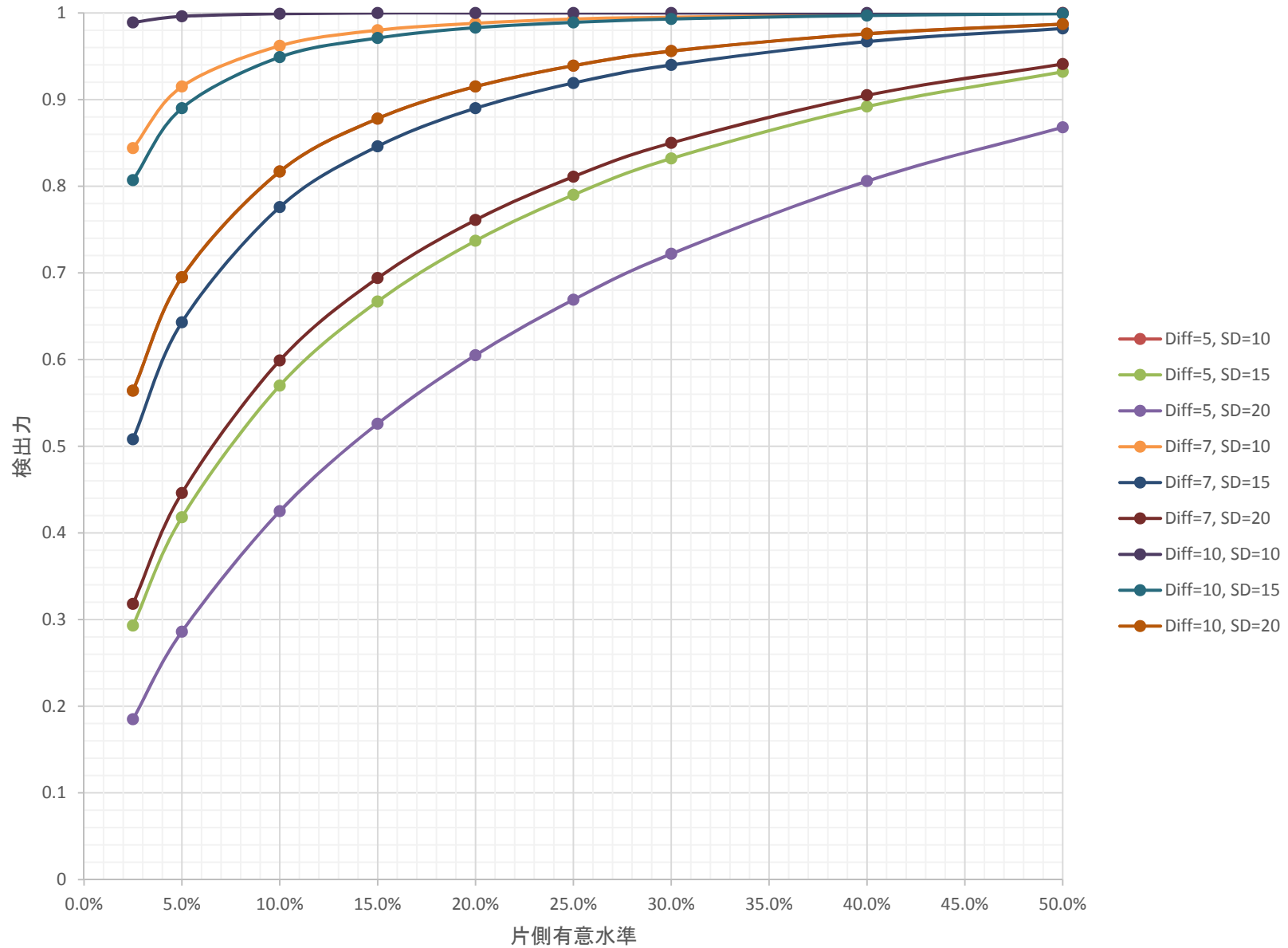
仮想ケースの参考資料(3/5)

体重とxの変化量



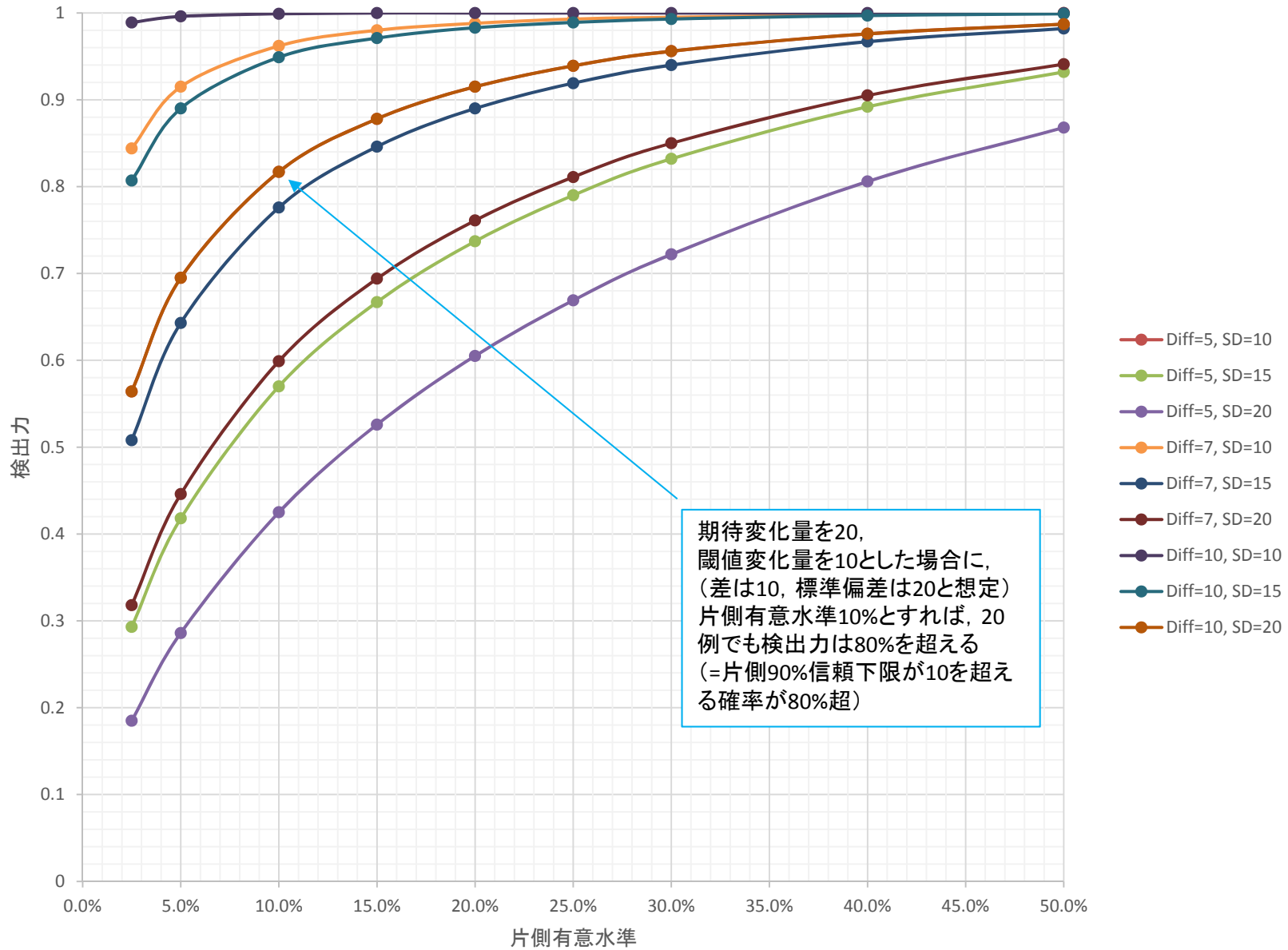
仮想ケースの参考資料(4/5)

片側有意水準と検出力(n=20)



仮想ケースの参考資料(5/5)

片側有意水準と検出力(n=20)



ディスカッション結果(1/2)

(各グループの意見を統合)

- 閾値
 - 設定: 良い結果を目指して20を設定, 選択基準で絞ることも検討
 - レジストリデータも考慮する。ガイドライン, KOLの意見も考慮。
 - 登録集団やアウトカムの影響を数値実験で性能評価して決める
 - 良い点: 基準が明確(規制当局とも合意しやすい)
 - 課題: 固定することで, ばらつきや対象集団の違いが考慮されない
 - 課題: 比較可能性をどう担保する? → 複数の閾値設定
 - 複数の集団で閾値を決めておき, その重みづけで判断した経験あり
- 外部対照
 - 良い点: 疾患対象をあるがままカバーできる
 - 今回の例数だとマッチングは難しい
 - 課題: 症例数設計はどうする? Makuch and Simon (1980) その他
 - 課題: 傾向スコアのモデリング, 傾向スコアの意味は?

ディスカッション結果(2/2) (各グループの意見を統合)

- Estimand
 - 対象集団をどこにおくかに依存
(それによって用いるエビデンス・方法論も異なる)
- 閾値・外部対照以外のフレームワーク
 - 自己対照, 推定精度ベースで評価(その他の方法も)
- データの質
 - CINのなかで研究班あり。GCP renovationの動きもあり。

ディスカッション2

ディスカッション2の目的

- ディスカッション2では、単群試験による薬効評価が許容されうる状況について、一般化することを目的とする

ディスカッション2の論点

- 2016年に開催されたEMAとESMOのワークショップ (EMA workshop on single-arm trials in cancer drug evaluation) でEMAから挙げられた3つの問題提起について議論する
 1. 無作為化比較試験が実施不可能な状況について、系統的に整理することはできないのか？
 2. ICH-E10において「治療効果が劇的」の場合に外部対照デザインが使用できると記載があるが、劇的の基準は何か？
 3. 試験デザインの計画において、単群で実施することにより生じる結果のバイアスを最小化する方策

論点1のディスカッション方法

- ディスカッション方法とし、臨床試験デザインに関するキーワードが書かれた紙を皆さんの前に配布済み
- キーワードを組み合わせて、無作為化比較試験が実施不可能な状況について系統的に整理をすることが目的
- キーワードの組み合わせについて、どういう場合には許容できそうか、できなそうかについて発言をお願いしたい

論点1のキーワード一覧

- プラセボ群の実施可能性なし
- 既存治療薬が存在しない
- 仮説検定の検出力を高く担保可能な例数の組入れが不可能
- ランダム化中止試験
- 希少疾病
- 小児
- がん
- 客観的評価指標
- 主観的評価指標
- 公表文献
- レジストリーデータ
- ガイドラインへの記載
- 評価指標に対する学会のコンセンサス
- 評価指標のバリデーション
- 学会又は患者の要望
- ランダム化対照試験の患者レベルのデータ
- 臨床的平衡 (Clinical equipoise)
- 分子標的薬/その併用
- 個別化医療
- 無作為化困難な治療 (遺伝子治療/細胞治療等)
- 劇的効果が認められている
- Mode of Action (薬効メカニズム)
- 内的妥当性/比較可能性
- 外的妥当性/一般化可能性
- 交絡
- 選択バイアス

論点2のディスカッション方法

- 劇的な効果の大きさについての共通認識を探ることが目的
- リスク差(割合の差)、ハザード比について、劇的と考える値を考えてください
- 劇的と考える値とその理由を説明してください
- どれくらいの値なら劇的と考えてよいのかについて皆で議論を行います

論点1のディスカッション結果 (グループ1)

- 希少疾病というだけでは、無作為化比較試験が実施できないとはいえない。小児といったより症例数が少ない条件も含まれる場合には、実施できない状況になりやすいだろう。
- 比較試験とした場合に生じる被験者組み入れの困難さ等、企業の実施可能性を大きく上回る場合で、かつ社会的に必要とされている薬の開発の場合には、やむを得ず無作為化比較試験を実施できないと判断できるかもしれない
- 自分の細胞から作った組織を移植する治療等で、医学的にプラセボでは有効性が全く期待できない場合には(結果として劇的な効果が期待される)、倫理的な理由も含め、無作為化比較試験は実施できないかもしれない
- レジストリーデータや公表文献、医学的常識等の観点から、ヒストリカルコントロールの結果が強く推測可能であり、開発薬剤の期待される結果が交絡因子等の影響を越えた差であること強く期待できる場合には、無作為化比較試験を実施しなくても有効性を評価できるかもしれない
 - 胃がん患者の手術後の再発率等
- 対照群をプラセボ群とするか、実薬対照又は標準治療群とするかで、無作為化比較試験の実施できない状況が変わってくると言う意見もあった
- その他として、仮に単群試験を計画した場合にヒストリカル対照するレジストリーデータの信頼性が担保されていれば、無作為化比較試験を実施しないという判断を後押しするかもしれない。また、主観的なエンドポイントでは単群試験での評価は難しいだろう。一方でハードエンドポイントだとしても、例えば死亡等は延命処置の具合で結果が変わるため、十分に注意する必要がある。

論点1のディスカッション結果 (グループ2)

- 無作為化試験が実施不可能な条件
(ORではなくAND条件)
 - 希少性(実施可能性)
 - 重症度(既存治療なし, 社会的要請含む)
 - 客観的評価による劇的な効果
 - 希少で重症だとしても, 客観的ではない評価によるものはバイアスの影響が大きい
ため判断が難しい
 - 重症+客観的評価による劇的な効果あり
→小規模なRCTが実施可能では?
 - 臨床試験データや疾患レジストリを活用しエビデンスを補強
- 実施可能性のみという場合も・・・
 - 手技の一部を変えるのみで差も小さくRCTをやるとなると現実的な期間に終わらない→単群試験
 - 医療機器の改良版評価のための単群試験

論点2のディスカッション結果 (グループ1)

- 割合の差の劇的な効果については、対照群の割合の結果に応じて以下のとおりであった。
 - 対照群:10%→開発薬剤群は50%
 - 10%程度しか有効性が期待できない場合に、半数は有効になるというのは劇的感がある
 - 対照群:50%→開発薬剤群は90%
 - 半数しか有効性が期待できない場合に、9割程度が有効というのは劇的感がある
 - 対照群:90%→開発薬剤群は99%
 - 9割程度が有効だとしても、薬剤によりほぼ100%有効になるというのは劇的感がある
 - 一方で、この場合には、劇的な差は考えられないという意見もあり
- ハザード比の劇的な効果については、対照群のハザードに依らず0.5以下という意見が多かった
 - がん領域では、0.6--0.7 程度のハザード比が期待できないと臨床試験の実施は勧められないという声明もあり、基本的にはそれ以下の数値が劇的なのだろう、という意見もあった

論点2のディスカッション結果) (グループ2) (1/3)

- 劇的効果の大きさ
 - 次ページに実際の数や人数を記載
- 割合のルール of 例
 - 閾値の2倍
 - 閾値と100%の間
- ハザード比のルール of 例
 - オッズ比0.5を下回ると結構インパクト大→HR=0.5
- 疾患・評価項目に大きく依存する
(それぞれ別々の疾患を想定している)

論点2のディスカッション結果 (グループ2) (2/3)

割合		リスク差
閾値	10%	X%
開発薬剤群	10% + X%	10, 40, 30, 40-45, 20, 30, 20, 10
閾値	60%	Y%
開発薬剤群	60% + Y%	30, 30, 30, 20, 30, 15, 10, 30
閾値	90%	Z%
開発薬剤群	90% + Z%	9, No, No - 10, 5, 10, 5, 5, 9

[XX]: XXは選んだ人数

時点Tにおける非イベント発生割合		ハザード比 劇的なし[1]
閾値	90%	0.10 [2]
開発薬剤群	99%	
閾値	90%	0.49 [4]
開発薬剤群	95%	
閾値	90%	0.79 [1]
開発薬剤群	92%	

論点2のディスカッション結果 (グループ2) (3/3)

[XX]: XXは選んだ人数

時点Tにおける非イベント発生割合		ハザード比
閾値	50%	0.15 [1]
開発薬剤群	90%	
閾値	50%	0.32 [3]
開発薬剤群	80%	
閾値	50%	0.51 [4]
開発薬剤群	70%	
閾値	50%	0.74
開発薬剤群	60%	

時点Tにおける非イベント発生割合		ハザード比
閾値	10%	0.52 [8] ただ、もう少し小さいほうが良い という意見も
開発薬剤群	30%	
閾値	10%	0.70
開発薬剤群	20%	
閾値	10%	0.82
開発薬剤群	15%	

パネルディスカッション

パネラー

- 浅野 (PMDA)
 - 兼清 (エーザイ)
 - 嘉田 (名古屋医療センター)
-

論点

- データの質
- 事前規定すべきポイント

データの質に関する議論

Q:データの質・目安は？

- (嘉田先生):論文はある程度、質が担保されていると想定。レジストリはその団体に対する信頼度も加味。できるだけ利用しようというスタンス
- (会場からコメント):RCTもレジストリも質は様々
- (兼清):PMDAの薬剤疫学研究実施のガイドラインが参考になるのでは？
- (会場からコメント):CINにおける疾患レジストリは、臨床開発でも利用可能な質の担保を目指している

Q:治療ガイドラインへの引用での目安は？

- (嘉田先生):GCP下でなくても引用する
- (会場からコメント):投資に対して出てきたものが少ないという現状改善を企図した21st Century Cures Actが成立した。この流れを踏まえると、利用するデータの質を、単に臨床試験と同程度まで上げるということはありません。
- (会場からコメント):GCP renovationにおける外部データ活用も同様の流れ

Q:そもそもデータの質とは？

- (会場からコメント):データの特徴(評価委方法,測定回数など)と質は分けて考えた方が良い。
- (会場からコメント):データを作成する際の手続きがしっかり守られていることがデータの質が担保されているということだと考えている。
- (会場からコメント):最近では、試験目的の評価に関する結果に強く影響を及ぼしうる項目について正しくデータが取られていれば、それをデータの質が担保されていると考えるという考えもある。

事前規定すべきポイントに関する議論(1/2)

Q: 以下の内容は事前に規定しておくべきと考えるが、過不足等あるか？

- 通常のRCTでも決めるもの (Estimand)
 - 対象集団
 - 評価項目
 - 盲検性の有無, 客観性
 - 解析手法
 - 解析対象集団
 - 主解析/sensitivity analyses
 - 症例数
- 閾値 (設定根拠含む)
- 対照群とするデータ
 - データソース, 選択・除外基準, データ数
 - 収集されている背景情報 (重要な背景情報の有無)
- 傾向スコアの推定方法 (評価含む) と利用した解析
 - 共変量の選択, 分布の評価方法
 - 処置群での/全集団での平均処置効果
 - 主解析/sensitivity analyses
- (嘉田先生): 傾向スコア以外はおおよそこれくらい。特に閾値の設定はしっかり議論

事前規定すべきポイントに関する議論(2/2)

Q: 傾向スコアに関する種々の条件を事前に決めることができるのか？

- (浅野): 傾向スコアの推定ができない場合や、得られた推定結果では、事前に規定していた共変量の調整がうまくいかない場合等あるように思う。上手くいかない場合にはどうするか等も事前に規定しておくのがよいのかもしれない。
- (兼清): ヒストリカルコントロールを用いた単群試験をよく行うFDA CDRH担当者の論文よりtwo-stage design for propensity scoreが提案されている
- (会場からのコメント): 主解析(ある1つの傾向スコア)だけでなく感度分析も重要なのでは？

参考文献

全般

- Eichler, H.-G., Bloechl-Daum, B., Bauer, P., Bretz, F., Brown, J., Hampson, L., ... Koenig, F. (2016). “Threshold-crossing”: A Useful Way to Establish the Counterfactual in Clinical Trials? *Clinical Pharmacology & Therapeutics*, 100(6), 699–712.
- Simon, R., Blumenthal, G., Rothenberg, M., Sommer, J., Roberts, S., Armstrong, D., ... Pazdur, R. (2015). The role of nonrandomized trials in the evaluation of oncology drugs. *Clinical Pharmacology & Therapeutics*, 97(5), 502–507.
- Cooper, B. S., Boni, M. F., Pan-ngum, W., Day, N. P. J., Horby, P. W., Olliaro, P., ... Whitehead, J. (2015). Evaluating Clinical Trial Designs for Investigational Treatments of Ebola Virus Disease. *PLOS Medicine*, 12(4), e1001815.

Propensity score関連

- D’Agostino, R. (1998). Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*, 2281, 2265–2281.
- Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46(3), 399–424.
- Yue, L. Q. (2012). Regulatory Considerations in the Design of Comparative Observational Studies Using Propensity Scores. *Journal of Biopharmaceutical Statistics*, 22(6), 1272–1279.
- Yue, L. Q., Lu, N., & Xu, Y. (2014). Designing Premarket Observational Comparative Studies Using Existing Data as Controls: Challenges and Opportunities. *Journal of Biopharmaceutical Statistics*, 24(5), 994–1010.
- Li, H., Mukhi, V., Lu, N., Xu, Y.-L., & Yue, L. Q. (2016). A Note on Good Practice of Objective Propensity Score Design for Premarket Nonrandomized Medical Device Studies with an Example. *Statistics in Biopharmaceutical Research*, 8(3), 282–286.
- Yue, L. Q., Campbell, G., Lu, N., Xu, Y., & Zuckerman, B. (2016). Utilizing national and international registries to enhance pre-market medical device regulatory evaluation. *Journal of Biopharmaceutical Statistics*, 26(6), 1136–1145.

参考文献

Sample size in trials using historical control 関連

- Song Zhang, Jing Cao, & Ahn, C. (2010). Calculating sample size in trials using historical controls. *Clinical Trials*, 7(4), 343–353.
- Makuch, R. W., & Simon, R. M. (1980). Sample size considerations for non-randomized comparative studies. *Journal of Chronic Diseases*, 33(3), 175–81.

Randomized registry trial 関連

- Fröbert, O., Lagerqvist, B., Gudnason, T., Thuesen, L., Svensson, R., Olivecrona, G. K., & James, S. K. (2010). Thrombus Aspiration in ST-Elevation myocardial infarction in Scandinavia (TASTE trial). A multicenter, prospective, randomized, controlled clinical registry trial based on the Swedish angiography and angioplasty registry (SCAAR) platform. Study design and. *American Heart Journal*, 160(6), 1042–1048.
- Fröbert, O., Lagerqvist, B., Olivecrona, G. K., Omerovic, E., Gudnason, T., Maeng, M., ... James, S. K. (2013). Thrombus Aspiration during ST-Segment Elevation Myocardial Infarction. *New England Journal of Medicine*, 369(17), 1587–1597.
- Lauer, M. S., & D'Agostino, R. B. (2013). The Randomized Registry Trial — The Next Disruptive Technology in Clinical Research? *New England Journal of Medicine*, 369(17), 1579–1581.