

RMST: Is it useful in non-inferiority trials ?

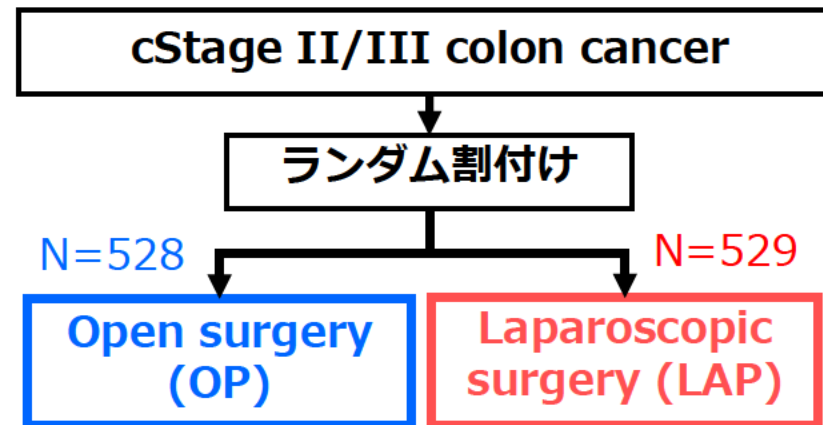
国立がん研究センター 研究支援センター 生物統計部
/ 東病院 臨床研究支援部門 生物統計室

野村 尚吾

※ 公開用にconfidentialな情報を省いています

前提知識の共有

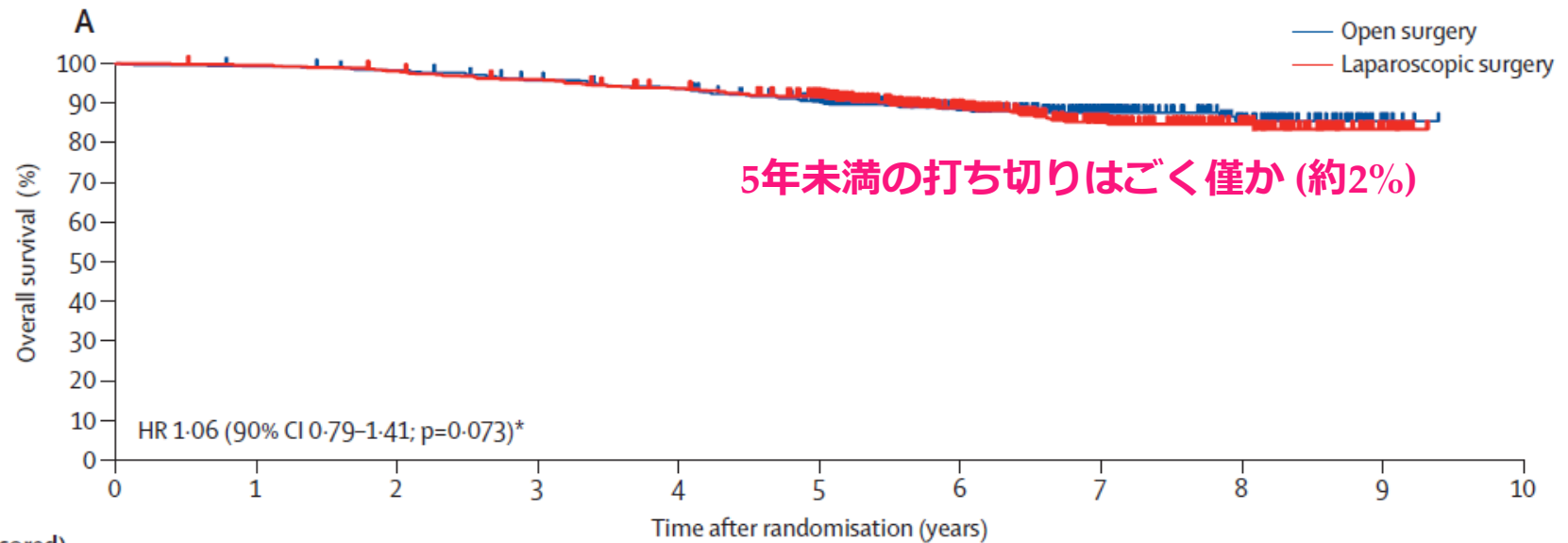
- 周術期を対象とした、がん領域の非劣性試験
 - 縮小手術, 低侵襲手術の意義を見る試験
 - 術後補助療法の期間短縮の意義を見る試験 など
- 例 : JCOG0404



primary: OS
片側 $\alpha=5\%$, power=80%

■ デザイン・解析手法の特徴

- HRを主解析とすることが一般的
- イベントを一定の数だけ集めてから解析
- **事前に定めた追跡期間までloss-to-followが少ない状態にすることが大切**
- 例：JCOG0404（追跡期間5年）



Number at risk (censored)

	0	1	2	3	4	5	6	7	8	9	10
Open surgery	528 (0)	523 (1)	516 (2)	499 (4)	486 (2)	442 (27)	291 (143)	164 (125)	71 (91)	12 (59)	0 (12)
Laparoscopic surgery	529 (0)	526 (1)	517 (1)	504 (2)	487 (5)	458 (19)	299 (147)	163 (127)	75 (87)	11 (63)	0 (11)

HRの問題点

- 予後良好集団では必要イベント数（検出力）確保のため、試験が大規模になってしまう
- 少数のイベントが解析結果に大きな影響力を持つ
 - 早期死亡例が偶然に一方の群に偏るだけで結果が変わるかもしれない
 - 選択・除外基準を絞る工夫（なるべく登録させないような工夫）が必要
 - 実際に登録してみて早期死亡のリスクが高いことが判明する場合がある（例：ランダム化後の術後病理診断から明らかに予後不良）
 - イベントが少数だと比例ハザード性の仮定が崩れやすいかもしれない

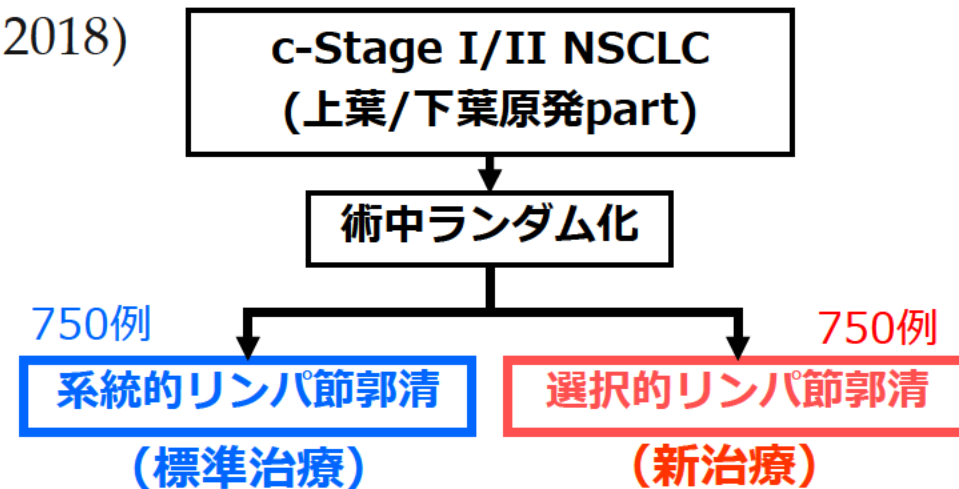


- 外科領域の臨床試験における問題点 @JCOG
 - HRを主解析に用いる非劣性試験を題材に説明します
- 非劣性試験でRMSTは有用か？
 - RMSTのpros/cons
 - JCOG試験での活用事例
- 優越性試験におけるRMST
 - おそらくJPMAの議論から漏れていること

JCOG1413

■ 計画時の仮説設定 (Hishida, JJCO 2018)

- %5yOS: 70% vs 70%
- 非劣性マージン: %5yOSで5.5%
 - HRに換算すると1.23
- 片側 $\alpha=5\%$, 検出力80%
- 登録5.5年, 追跡5年



primary: OS

計画時の必要イベント数: 579

モニタリング時に判明したこと

■ 登録ペース良好

■ 予後良好

—

A large black rectangular redaction box covers the content of the slide under the '予後良好' section. A small horizontal line is visible to the left of the redaction.

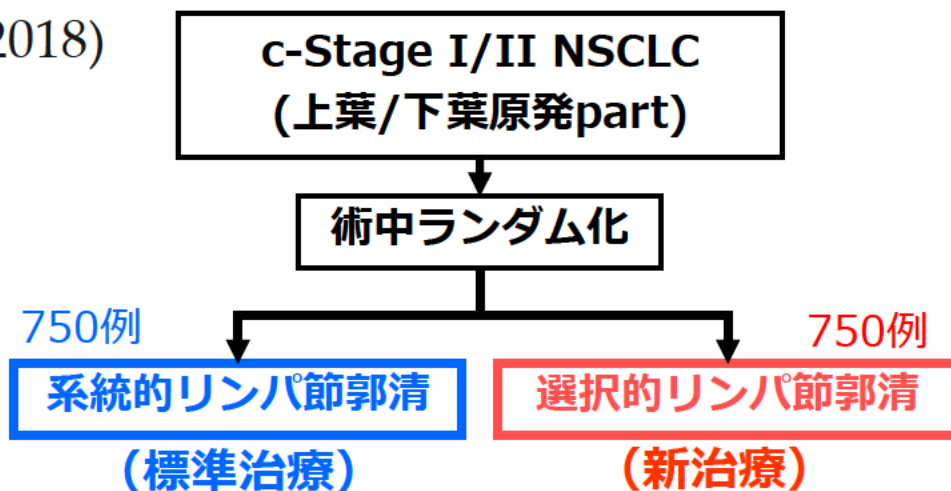
JCOG1413 (改訂後)

■ 計画時の仮説設定 (Hishida, JJCO2018)

- %5yOS: 70% vs 70%
- 非劣性マージン: %5yOSで5.5%
 - HRに換算すると1.23
- 片側 $\alpha=5\%$, 検出力80%
- 登録5.5年, 追跡5年

■ 改訂後の仮説設定

[Redacted text]



primary: OS

改訂前の必要イベント数: 579

改訂後の必要イベント数: [Redacted]

予後良好な集団における周術期の非劣性試験

- JCOG0404: cStage II/III colon cancer (5yOS=82%)
 - Open surgery vs Laparoscopic surgery, N=1050
- JCOG0802: cStage I small peripheral NSCLC (5yOS=90%)
 - Lobectomy vs limited resection (segmentectomy), N=1100
- JCOG0912: cStage I gastric cancer (5yOS=90%)
 - Open surgery vs Laparoscopic surgery, N=920
- JCOG1601 cStage I/II tongue cancer (5yOS=85%)
 - Prophylactic neck dissection vs partial glossectomy alone, N=440

同様の事例: JCOG0404

■ サンプルサイズ設定

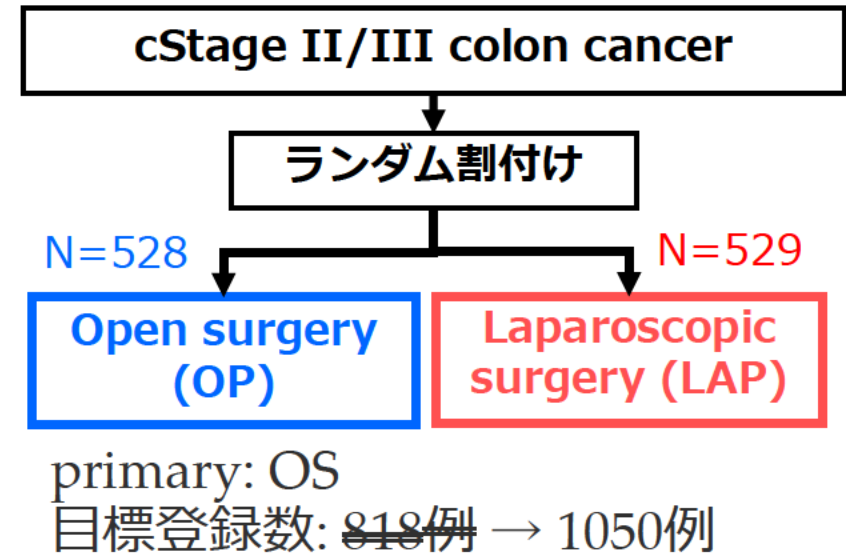
- **5y%OS: 75% vs 75% ($\Delta=7.5\%$)**
 - HRに換算すると1.366
- **必要イベント数: 254**
- 登録3年・追跡5年

■ 盲検下のモニタリング結果

- イベント数が極端に少ないことが判明
- 他試験の成績から予後の見込みを上方修正

■ 改訂後の想定（必要イベント数は不変）

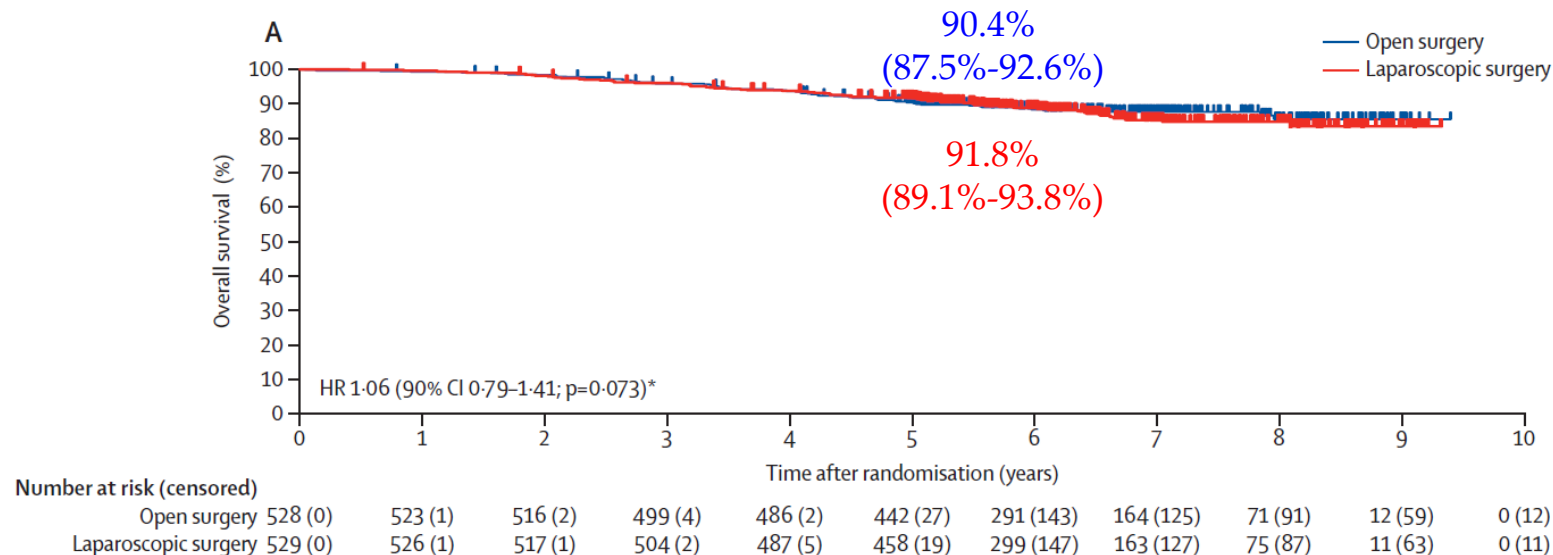
- **5y%OS: 82% vs 82% ($\Delta=5.7\%$) ※HRに換算したマージンは1.366のまま**
- **登録を3年から4.5年に延長**
- **必要イベント数: 254のまま、目標登録数を1050に上方修正**



■ 登録終了後5年間追跡したが128イベントしか集積されず

- 目標イベント数254の約半数 (大幅な検出力不足)
- しかし, これ以上追跡を継続するのは困難だった・・・
 - 登録開始から10年以上が経過してしまう
 - 5年未満の打ち切りは僅か。5年生存割合の比較結果が知りたい!

■ 検出力不足だが, 主解析を実施



- 曲線は重なって見えるが, **主解析の結果はnegative**

想定より予後良好だった場合の対応

- 試験を中止する判断には至らないだろう
 - 想定より予後良好だから、無効中止には至らない
 - 僅かなイベント数では有効中止に至る確率は極めて小さい
- 盲検下のサンプルサイズ再設定が望ましい
 - 変更しないパラメータ： 追跡期間
 - 変更するパラメータ
 - 登録期間と想定する予後: 進捗状況から見直す場面あり
 - **非劣性マージン (Δ): 計画時の値をそのまま使うべき?**

非劣性試験: 盲検下のサンプルサイズ再設定

- 非劣性マージンについて -

■ 周術期の試験では「臨床的に許容しうる最大の差」から非劣性マージンを設定する場面が多い

- FDAガイダンスの「M2」に相当
- M1も重要だが、「手術なし」の成績がない/古い状況がある
- JCOG0404/1413では%5yOSの「差」からM2を設定

■ 非劣性マージンの変更：2つの方針

- JCOG0404: 予後が良くなってもHR=1.366のまま（**比を優先**）
 - %5yOSのスケールでは計画時より小さな値を見込むことに
 - **帰無仮説が変わってしまうので比を優先したい統計家が多いだろう・・・**
- JCOG1413: 予後が良くなっても $\Delta\%5yOS = -5.5\%$ のまま（**差を優先**）
 - HRスケールでは計画時より大きな値を見込むことに
- **試験治療のメカニズムによって考え方が変わる**

ここまでのまとめ

- 外科領域の非劣性試験における現状を紹介
 - 予後が良く, 必要イベント数に (遠く) 達しない
- それに伴う試験計画の変更 (HRで解析することを前提)
 - 事後的に想定する予後を上方修正する改訂が必要になる
 - 非劣性マージンをどのように見直すか? が1つの論点
 - 「差」と「比」のどちらを優先するか
- そもそもHRに固執する必要があるのだろうか?

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Alternatives to Hazard Ratios for Comparing the Efficacy or Safety of Therapies in Noninferiority Studies

Hajime Uno, PhD*; Janet Wittes, PhD*; Haoda Fu, PhD*; Scott D. Solomon, MD; Brian Claggett, PhD; Lu Tian, ScD; Tianxi Cai, ScD; Marc A. Pfeffer, MD, PhD; Scott R. Evans, PhD; and Lee-Jen Wei, PhD



- 外科領域の臨床試験における問題点 @JCOG
 - HRを主解析に用いる非劣性試験を題材に説明します
- 非劣性試験でRMSTは有用か？
 - RMSTのpros/cons
 - JCOG試験での活用事例
- 優越性試験におけるRMST
 - おそらくJPMAの議論から漏れていること

HR以外の指標 (非劣性試験の場合)

※ 簡単のため, 差に基づく指標だけを掲載している

- $\Delta\%$ t-year milestone survival: $\Delta\hat{S}(t) = \hat{S}_1(t) - \hat{S}_0(t)$
- Δ MST (percentile): $\Delta\hat{S}^{-1}(0.5) = \hat{S}_1^{-1}(0.5) - \hat{S}_0^{-1}(0.5)$
- **ARMST**: $\Delta \int_0^\tau S(u)du = \int_0^\tau S_1(u)du - \int_0^\tau S_0(u)du$
 - 対象者が τ まで無イベントである期間の期待値 (の差)
- Δ RMTL: $\Delta \int_0^\tau 1 - S(u)du$
 - τ から RMST を引いたもの (の差)
 - イベント数が少ない場合, RMTL の比は HR の近似
- 注: 以降では, 主解析の変更以外のアプローチは考えない
 - α/β を緩める
 - 登録期間/追跡期間を延長
 - RFS/DFS といった代替指標に primary endpoint を変更

Δ RMSTのpros/cons @非劣性試験

- pros -

Uno, et al. Ann Intern Med 2015.より一部引用 / Sciria, et al. N Engl J Med 2013. Fig1A

- 曲線全体を比較できる ($\Delta\%$ 5yOSや Δ MSTに対する優位性)
- 「モデルフリー」というメリット (HRに対する優位性)
- 解釈のしやすさ (HRに対する優位性)

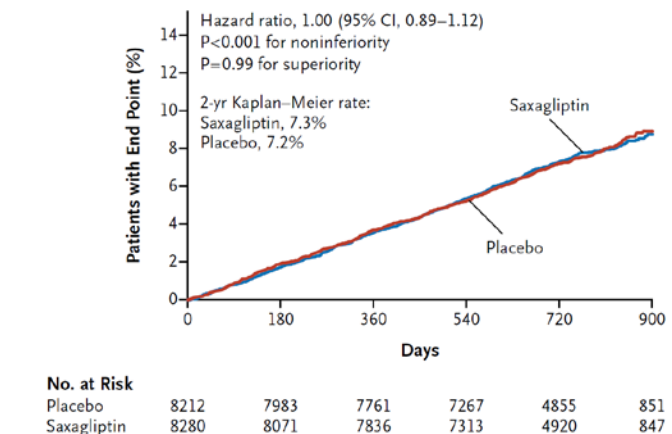
– 例) 2型糖尿病のplacebo vs saxagliptin (心臓血管イベントに着目)

- HR=1.00 [95%CI, 0.89 to 1.12]
- $\tau = 900$ 日としたとき, 両群ともRMST = 860日
- Δ RMST = 0 days [95%CI, -5 to 4]

– RMSTの解釈:

「心臓血管イベントがない日数」は、
投与から900日経過した段階で両群とも860日。
saxagliptinはplaceboより最悪で5日短くなる。

A Primary End Point



Δ RMSTのpros/cons @非劣性試験

- pros [つづき] -

- 一定期間, きちんと追跡して解析するのが定石だから, 臨床家が納得できる τ を決めやすい状況が多いだろう
 - 一方で進行がん領域の試験では τ の設定は難しい
- 漸近相対効率の比較 (Tian, Biometrics 2018)
 - PH: 大差なし
 - non-PH
 - crossing hazards, diminished effect: RMSTの方が良い
 - delayed effect: HRの方が良い

※ 上記は限られたシナリオでの検討である点に注意
(実際の臨床試験データに近いワイブル分布に基づく検討)

Δ RMSTのpros/cons @非劣性試験

- cons -

Uno, et al. Ann Intern Med 2015.より一部引用

- **現状, がん領域のcommunityで広く受け入れられていない**
 - ステークホルダーを巻き込んで慎重に議論を進める必要がある
 - 医学雑誌・診療ガイドラインの各種委員に「適切なエビデンス」と解釈させることが重要ではないか
- **参照値とセットにした評価が不可欠 (HRでは不要)**
 - 「何に対する Δ RMST = -5日なのか？」を評価することが重要
 - 対照群のRMSTが30日なのと365日なのでは臨床的な意味合いが異なる
- **予後良好な集団を登録するようなバイアスが生じないか？**
 - 予後良好例を選択的に登録することで Δ RMSTが小さくなり、非劣性が証明されやすくなるかもしれない
(※ RMSTに限った問題ではない)

Δ RMSTのpros/cons @非劣性試験

- cons [つづき] -

■ τ の設定

- 再掲： 周術期の試験では τ （の下限）は決めやすい
- τ 以降の生存曲線を見捨てて良いかについては議論があるだろう
- 「追跡期間」と「登録期間 + 追跡期間*」のどちらが良いか？
* 正確には登録期間 + 追跡期間内における死亡例の最長OS
 - 登録状況・登録/追跡の構成割合・打ち切りの分布等によって性能は変わるだろう
 - その性能は十分に調べられていない

■ 中間解析

- どの解析時点でも同じ τ を用いるならば Δ RMSTのZ統計量は独立増分
- (OSが) τ に達した症例が何割いれば妥当な推測なのか？
 - 生存曲線の右端（横軸）が τ に達しないとき、どうやって解析するのか？



- 外科領域の臨床試験における問題点 @JCOG
 - HRを主解析に用いる非劣性試験を題材に説明します
- 非劣性試験でRMSTは有用か？
 - RMSTのpros/cons
 - JCOG試験での活用事例
- 優越性試験におけるRMST
 - おそらくJPMAの議論から漏れていること

JCOG1413: 主解析計画の追記 ()

- 観察イベントが目標の70%を下回る場合に Δ RMSTに変更
 - %5yOSが を上回る状況に相当
 - 追跡期間から $\tau = 5$ 年と設定
- 数値実験で動作特性を確認

※ 改訂後の研究計画書を参考に野村が実施

■ 条件

- サンプルサイズ：各群■例（改訂後の目標登録数）
- 登録■年、追跡5年
- 生存時間に仮定する分布：指数分布
- 有意水準：片側 $\alpha=5\%$ （中間解析は考慮しない）

■ 評価

- 2万回繰り返し
- HR/ Δ RMSTの第1種の過誤率・検出力を計算

JCOG1413数値実験

- 結果 -

両群を併合 した%5yOS	期待イベント数	必要イベント数 に占める割合	HR	Δ RMST
		1	89.6%	83.4%
		0.9	86.5%	86.5%
		0.8	83.2%	89.6%
		0.7	78.4%	92.7%
		0.6	72.8%	95.5%
		0.5	65.7%	97.8%
		0.4	57.5%	99.2%
		0.3	48.4%	99.8%
		0.2	36.9%	100.0%
		0.1	24.7%	100.0%

なぜ Δ RMSTの検出力が高いのか？

- 常に Δ RMSTの非劣性マージンが-0. [] 年だから
- Δ RMST = -0. [] 年をHRや $\Delta\%$ 5yOSのスケールに換算するとどうなるだろうか？

Δ RMSTの非劣性マージン = -0.■■■年とする場合

- $\Delta\%5yOS$ ・HRとの対応 -

■ HR

- 段々と緩いマージンに

■ $\Delta\%5yOS$

- ほとんど同じマージン

両群を併合した%5yOS	マージン		
	$\Delta\%5yOS$	HR	Δ RMST
■■■	-5.5%	■■■	-0.■■■ (固定)
■■■	-5.5%	■■■	
■■■	-5.6%	■■■	
■■■	-5.6%	■■■	
■■■	-5.6%	■■■	
■■■	-5.7%	■■■	
■■■	-5.7%	■■■	
■■■	-5.7%	■■■	
■■■	-5.8%	■■■	
■■■	-5.8%	■■■	

Δ RMSTでマージンを固定 \equiv $\Delta\%5yOS$ でマージンを固定

予後良好になるにつれ、HRのマージンはかなり大きな値になってしまう

(注意：HRのマージンが大きな値でも臨床的に緩い設定とは限らない！)

Δ RMSTの非劣性マージン = -0. ■■■年とする場合

- 小括 -

- 想定より予後良好だと Δ RMSTはHRより検出力が高い
 - 計画時より「緩い」HRのマージンで検定することに対応する点を批判する臨床家・統計家は一定数いるだろう
- では、常にHRのマージンを■■■に固定したらどうなるか？

HRの非劣性マージン = ■■とする場合

- $\Delta\%5yOS \cdot \Delta RMST$ との対応 -

- $\Delta RMST / \Delta\%5yOS$
- 段々と厳しいマージンに

「■■%に対する-5.5%は許容」
でも
「■■%に対する■■%は許容しない」
比を優先する考え方は
状況次第では厳し過ぎる設定？

両群を併合した%5yOS	マージン		
	$\Delta\%5yOS$	HR	$\Delta RMST$
■■■	-5.5%	■■■ (固定)	■■■
	-5.0%		
	-4.5%		
	-3.9%		
	-3.4%		
	-2.8%		
	-2.3%		
	-1.7%		
	-1.2%		
	-0.6%		

HRの非劣性マージン = ■■とする場合

- 検出力 -

両群を併合した%5yOS	期待イベント数	必要イベント数に占める割合	HR	Δ RMST
■■	■■	1	89.6%	83.4%
■■	■■	0.9	86.5%	79.8%
■■	■■	0.8	83.2%	75.6%
■■	■■	0.7	78.4%	70.6%
■■	■■	0.6	72.8%	64.8%
■■	■■	0.5	65.7%	58.2%
■■	■■	0.4	57.5%	50.5%
■■	■■	0.3	48.4%	41.8%
■■	■■	0.2	36.9%	32.7%
■■	■■	0.1	24.7%	22.4%

Δ RMSTは総じて検出力が低い

HRの非劣性マージン = ■■■ とする場合

- 検出力 (Δ RMSTの τ を動かしたとき) -

両群を併合 した%5yOS	期待イベント数	必要イベント数 に占める割合	HR	Δ RMST		
				$\tau=3y$	$\tau=5y$	$\tau=8y$
■■■	■■■	1	89.6%	66.7%	83.4%	93.6%
		0.9	86.5%	62.5%	79.8%	91.5%
		0.8	83.2%	58.3%	75.6%	88.7%
		0.7	78.4%	53.2%	70.6%	84.6%
		0.6	72.8%	48.2%	64.8%	79.8%
		0.5	65.7%	42.6%	58.2%	73.5%
		0.4	57.5%	36.7%	50.5%	65.1%
		0.3	48.4%	30.5%	41.8%	54.8%
		0.2	36.9%	24.4%	32.7%	43.2%
		0.1	24.7%	17.3%	22.4%	28.4%

$\tau = 8y$ のとき, HRより検出力が高い

結果の整理

- 「差」を優先するマージン設定
 - 予後良好になるにつれ、緩いHRのマージンで検定することに対応
 - Δ RMSTの方が検出力が高い
- 「比」を優先するマージン設定
 - 予後良好になるにつれ、厳しい Δ RMSTのマージンで検定することに対応
 - 検出力は τ の設定次第

予後・マージンの対応関係を計画時に整理することが重要ではないか？（以下は仮想例）

両群を併合した%5yOS	マージン		
	$\Delta\%5yOS$	HR	$\Delta RMST$
	-5.5%	■	-■
	HR=■を固定 (比を優先, 主解析はHR)		
	$\Delta RMST$ =■を固定 (差を優先, 主解析は $\Delta RMST$)		

この方針に沿って盲検下のSSRプランを事前に計画書やSAPで規定しておく

※ モニタリングや他試験の結果によっては、試験途中にこの対応関係を見直したい状況が生じるだろう

まとめ（まだ終わりません）

- 周術期の非劣性試験では計画時の想定より予後良好になる場面があるが、HRでは検出力が著しく低下しうる
- Δ RMSTの動作特性は非劣性マージンや τ の設定に依存
- 予後・非劣性マージンの関係を事前に整理することは一案



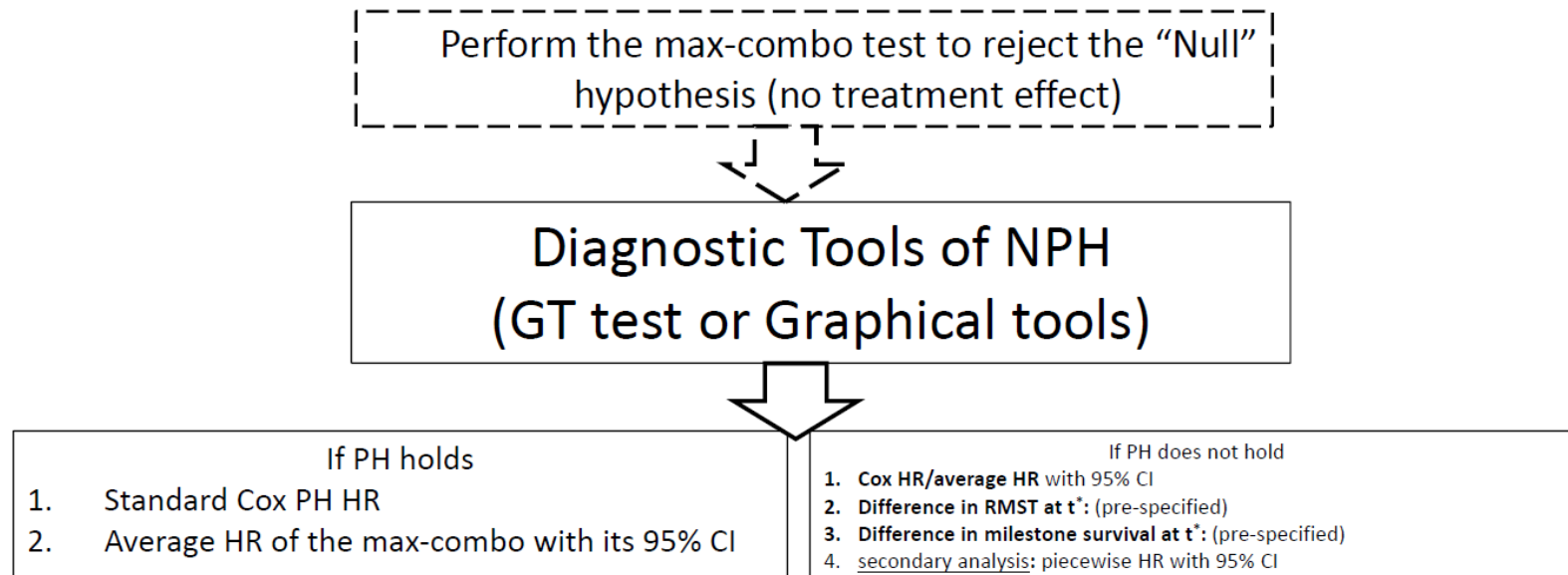
- 外科領域の臨床試験における問題点 @JCOG
 - HRを主解析に用いる非劣性試験を題材に説明します
- 非劣性試験でRMSTは有用か？
 - RMSTのpros/cons
 - JCOG試験での活用事例
- 優越性試験におけるRMST
 - おそらくJPMAの議論から漏れていること

優越性試験の主解析

- PHが妥当なら既存の手法(LRT/Cox)が良い
- non-PHで Δ RMSTを主解析とすべき？
 - non-PHにおける既存手法の問題
 - 効率低下
 - 解釈が難しい
 - PH/non-PHの判定は主観が伴う
 - 複数の指標を計算し、多面的に評価するのが現実的に思う
 - 1つのツールとして、RMSTは優れた手法

■ NPH working groupの提案

- 「**primary analysisでなく primary analysis procedureを事前規定**」
(Royston, et al. Stat Med 2011)



- 対応する推定の手法がないため, 気持ち悪さが残るが, MaxComboを主解析とするのは妙案だろう (私感)
 - もちろん, ステークホルダーを巻き込んだ議論が必要

おそらくJPMAの議論から漏れていること

- non-PHには「原因」があり、
その究明が極めて重要である
 - がん領域では多くの新規技術が臨床導入されている
 - 次世代シーケンサーに基づく遺伝子パネル
 - 免疫細胞の発現解析
 - cell free DNAの解析 (liquid biopsy) など
- 生存曲線の型に応じた「良い」解析アプローチが分かってても non-PHの原因が分かったわけではない
 - 事後的であっても、「なぜ、その曲線のタイプになったのか」
深く考察することが重要

生存曲線のタイプに応じた考察

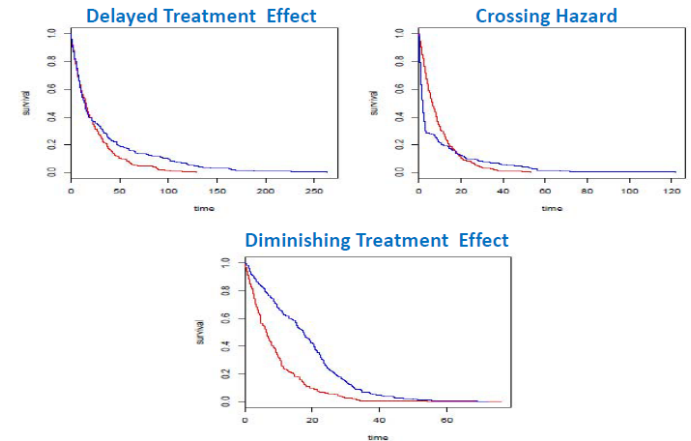
■ ここで注目する生存曲線のタイプ

- delayed scenario, crossing scenario
- diminishing scenario

■ delayed scenario, crossing scenario

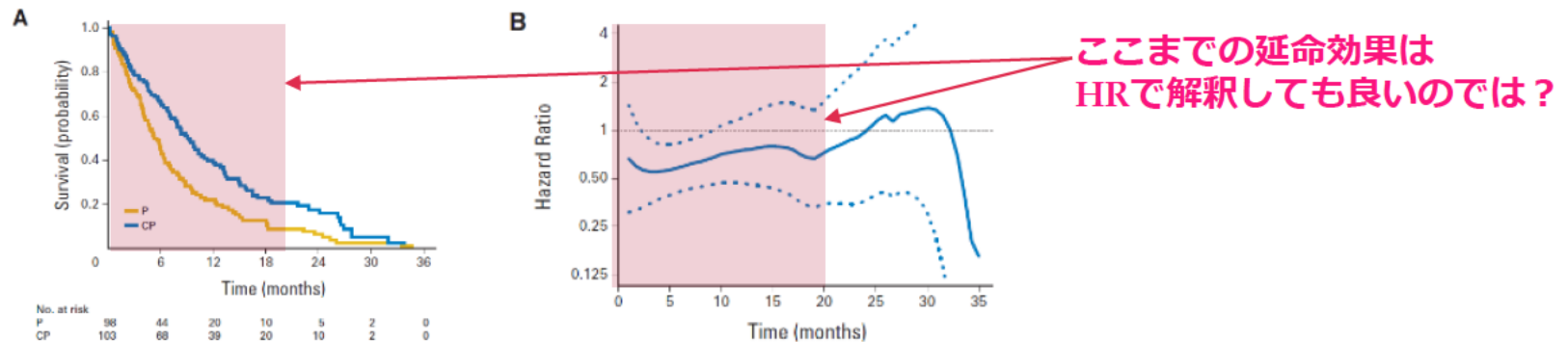
- 有効なサブセットの存在が強く懸念される
- 効果予測マーカーの探索が重要
 - 免疫CP阻害薬: PD-L1, tTMB-h, bTMB-h, microbiome
 - 分子標的薬: driver mutation, mutational type
 - その他の薬剤: multi-gene assay (e.g., OncotypeDX assay)

Different Types of NPH



■ diminishing scenario: 多くの進行がんに対する臨床試験

- 進行がんの優越性試験（特に医薬品開発を目的とした治験）では「一定期間（例えばMST付近まで），生存曲線が優っていることを(短い追跡期間で)証明すること」に関心があるのでは？
- 最終的には大部分の患者が亡くなってしまいうため追跡を継続してnon-PHの傾向が強くなるのは当たり前



- 長期追跡後はnon-PHが強く想定されるため, RMST等の指標を提示して総合的に延命効果や治癒例について考察すべきだろう



**RMST:
Is it useful in non-inferiority trials ?**

YES.

Especially, in peri-operative oncology settings.

