

JCOG臨床試験セミナー中級編

2014年10月11日

生存時間解析

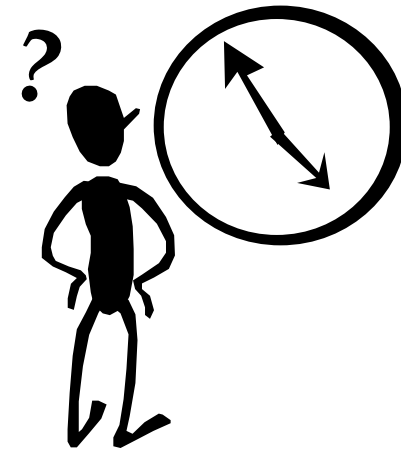
国立がん研究センター 研究支援センター
野村 尚吾

“3種の神器”では解釈が難しいケースがある

※“3種の神器”の詳細については、ICRwebの「生存時間解析」の講義をご覧ください

■ 生存時間解析の“3種の神器”

- **Kaplan-Meier法**
 - 生存時間データを要約する手法
- **ログランク検定**
 - 生存曲線を比較する手法
- **Cox回帰**
 - 共変量と生存時間をモデル化する手法



こういったケース？
その対処法とは？

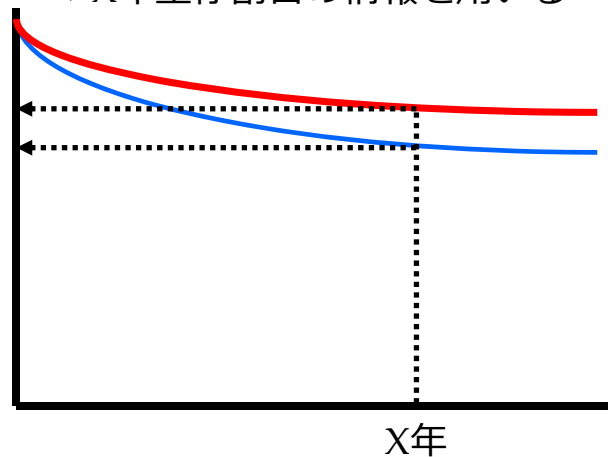
OUTLINE

- Kaplan-Meier法は万能か？
 - ハザード関数の紹介
- ログランク検定は万能か？
 - “delayed effect”の問題について
 - Fleming-Harrington検定の紹介
- Cox回帰は万能か？
 - ハザード比による要約は適切？

デザイン段階では指数分布を仮定することが一般的

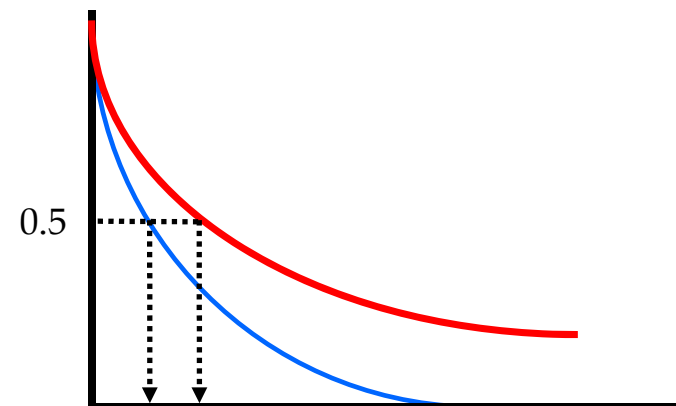
■ サンプルサイズ設定時にハザード比を見積もる場合

adjuvantなど、予後良好な場合
⇒ X年生存割合の情報を用いる



$$\text{ハザード比} = \frac{\log(\text{試験群の X年生存割合})}{\log(\text{対照群の X年生存割合})}$$

Stage IV など、予後不良の場合
⇒ MSTの情報を用いる

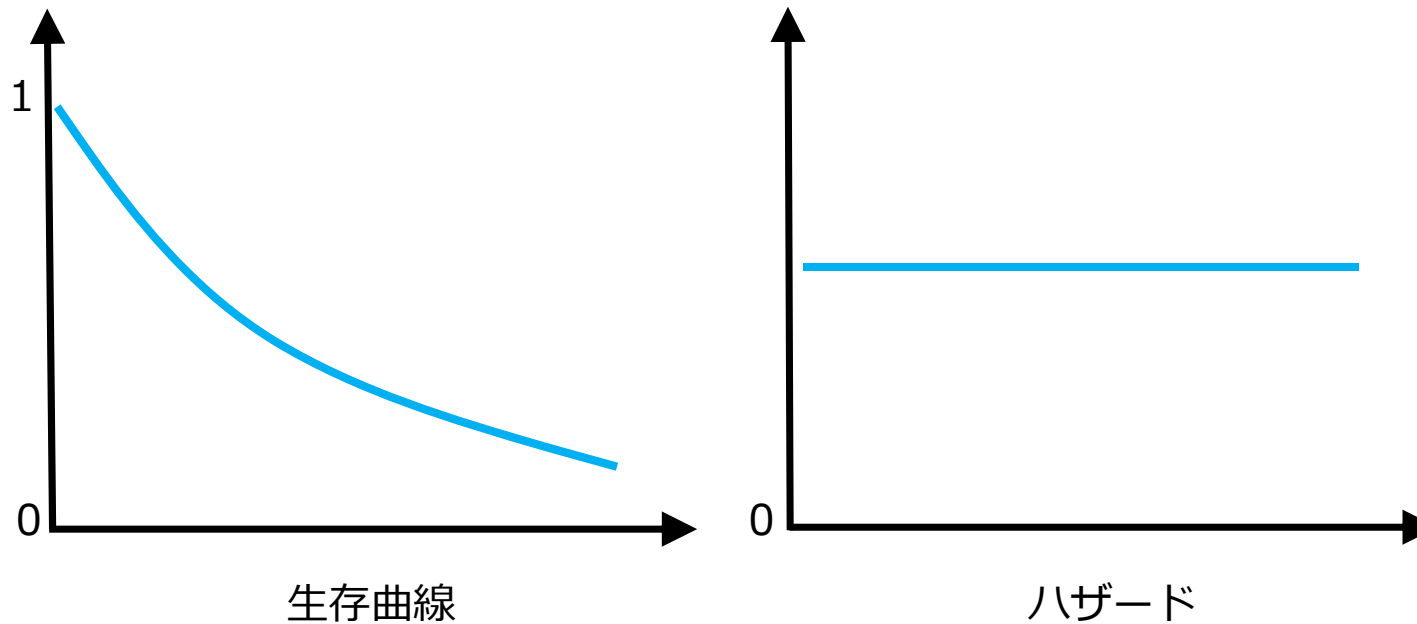


$$\text{ハザード比} = \frac{\text{対照群の MST}}{\text{試験群の MST}}$$

※ サンプルサイズ設定で通常用いられる「Schoenfeld & Richter の方法」も指数分布を仮定

指数分布

■ 全時点でハザードが一定と仮定した確率分布



— 「ハザードが一定」をどうやって確かめる？

ハザード

■ ある時点の瞬間イベント発生率

– 時点 t のat risk集団が時点 t で死亡する確率

- 単位: [イベント数/人年]
- 0より大きな値をとる

– Kaplan-Meier法に基づくハザードの推定

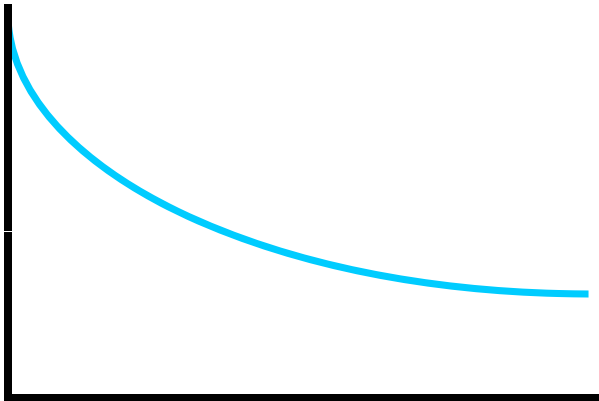
- イベント時点間ではハザードは一定と仮定
- 第 j イベントから第 $j+1$ イベントまでの期間 Δt のハザード
※ 第 j イベント発生時のリスク集合(#at risk)、イベント数(#event)

$$\text{ハザード} = \frac{\# \text{ event}}{\# \text{ at risk} \times \Delta t}$$

ハザードと生存曲線の関係

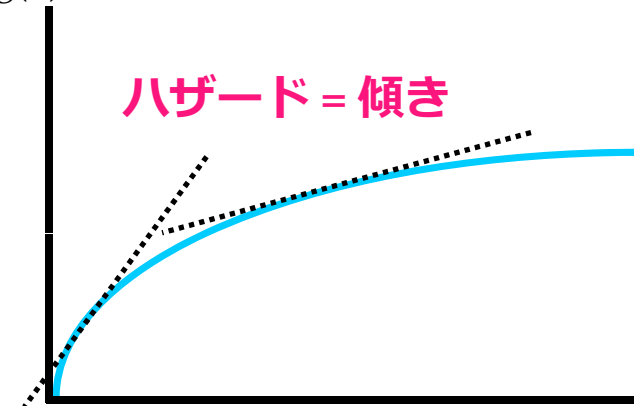
指数分布以外の場合：時点によってハザードの値が変わりうる

生存関数 S



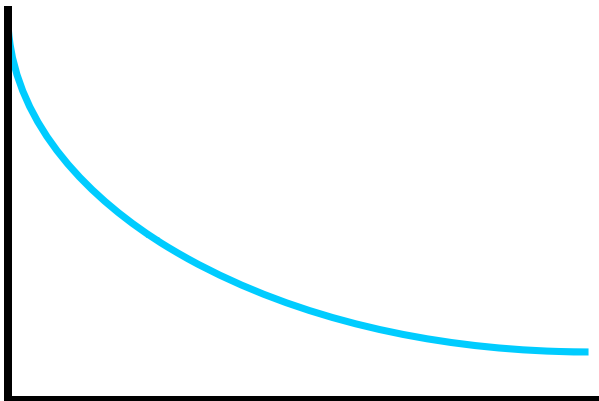
対数をとって
マイナス1倍

$-\log(S)$: 累積ハザード関数



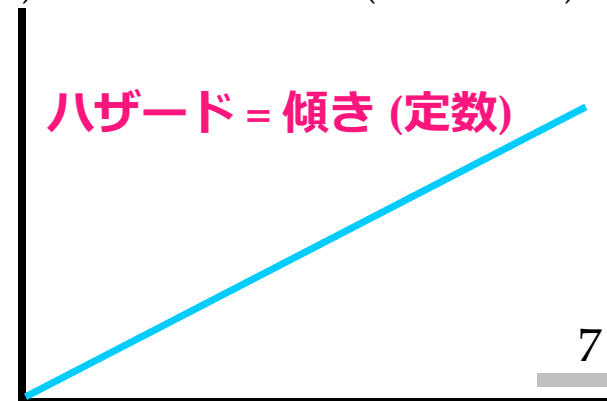
指数分布の場合：ハザードは常に一定

生存関数 S



対数をとって
マイナス1倍

$-\log(S)$: 累積ハザード関数(直線になる)



ハザード関数の利用について

Statistics in Clinical Cancer Research

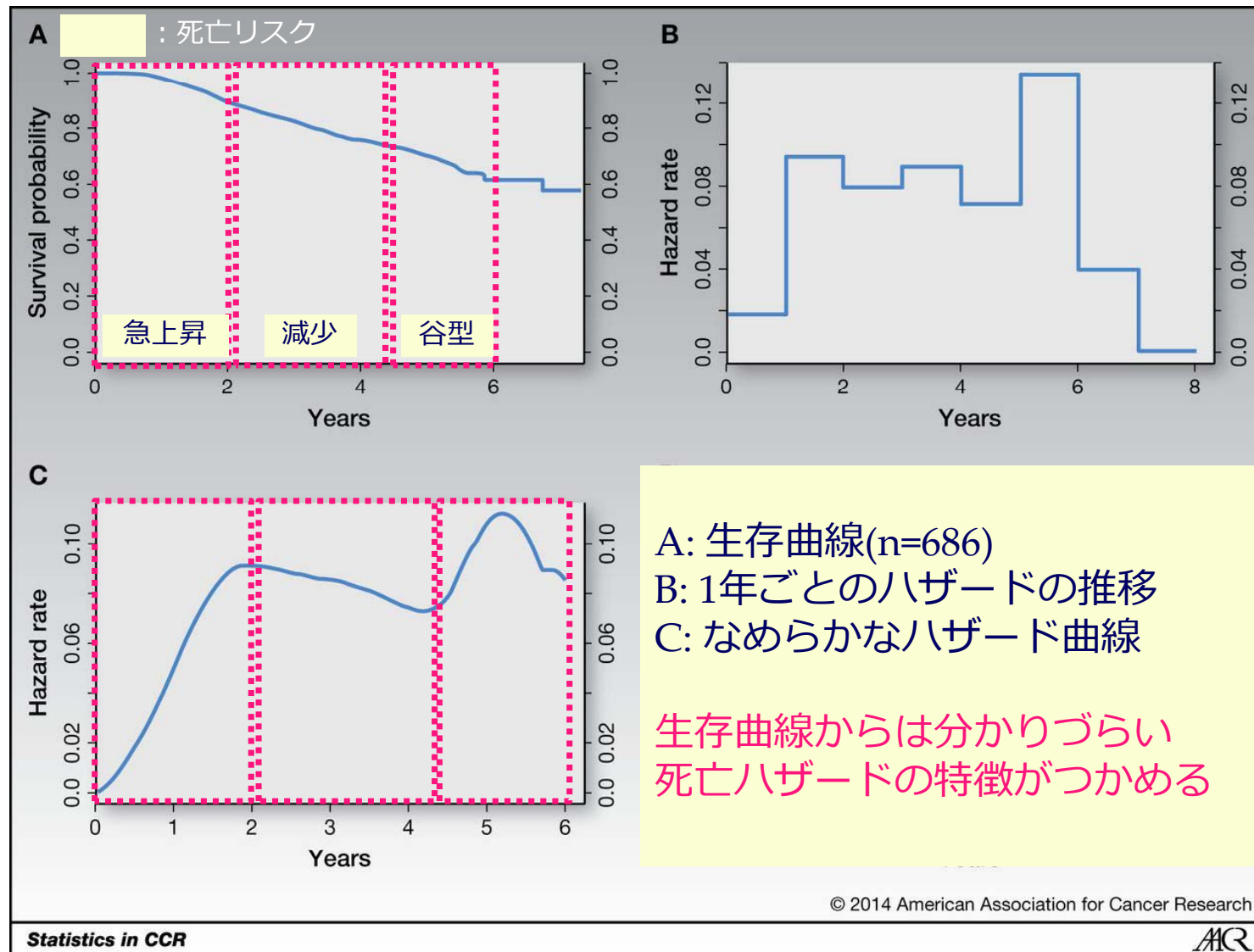
**Clinical
Cancer
Research**

Getting More Out of Survival Data by Using the Hazard Function

Kenneth R. Hess¹ and Victor A. Levin²

－ハザード関数についてまとめたレビュー論文

Hess K and Levin V. Clin Cancer Res 2014. Fig1より抜粋



- 
- Kaplan-Meier法で推定した生存曲線を見ても
ハザード関数の概要はつかめない
 - ハザードはKM曲線の追加情報として有用
 - リスクの経時的変化が検討可能
 - デザイン・解析時の仮定の妥当性が検討可能
 - 指数分布にしたがうか？
 - 比例ハザード性の仮定が正しいか？

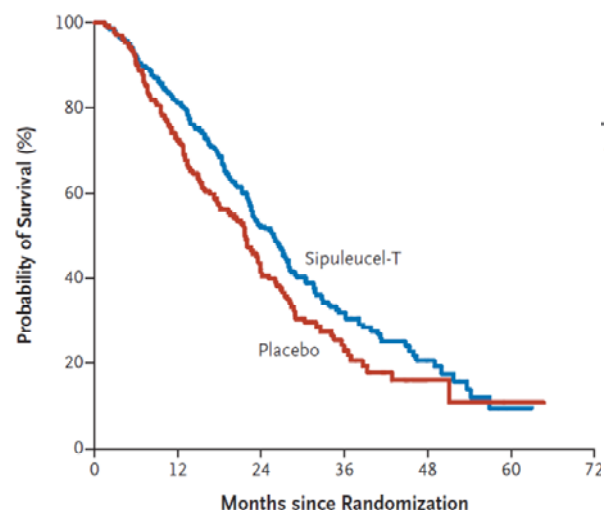
OUTLINE

- Kaplan-Meier法は万能か？
 - ハザード関数
- ログランク検定は万能か？
 - “delayed effect”の問題について
 - Fleming-Harrington検定の紹介
- Cox回帰は万能か？
 - ハザード比による要約は適切？

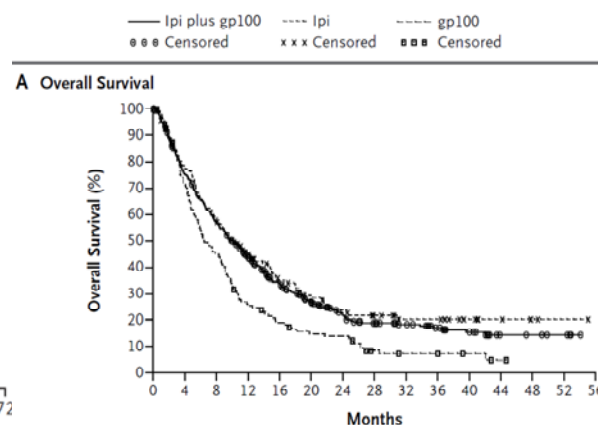
がん免疫療法の臨床試験

■ 治療効果が遅れて認められる場合がある

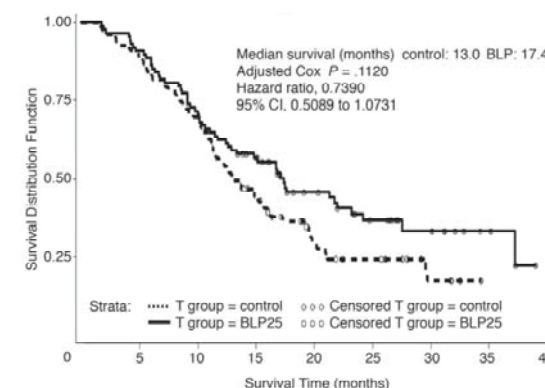
- 増悪後の腫瘍縮小 (Wolchok J, et al. Clin Cancer Res 2009など)
- 典型的な生存曲線



Kantoff P, et al. New Engl J Med 2010



Hodi F, et al. New Engl J Med 2010



Butts C, et al. J Clin Oncol 2005

“*delayed vaccine effect*”の問題点

■ 比例ハザード性が成立しないかもしれない

show no effect for the initial portion of the study. If the treatment is effective, separation of the curves may occur later in the study after the vaccine effect has become established. This may violate the proportional hazard assumptions necessary when applying Cox modeling and may necessitate an increase in sample size to provide sufficient power to test a statistical hypothesis.⁴

FDA draft guidance (2009). “Clinical Considerations for Therapeutic Cancer Vaccines”: p.11.

■ 観察期間後期に重みをおいた検定手法が有用

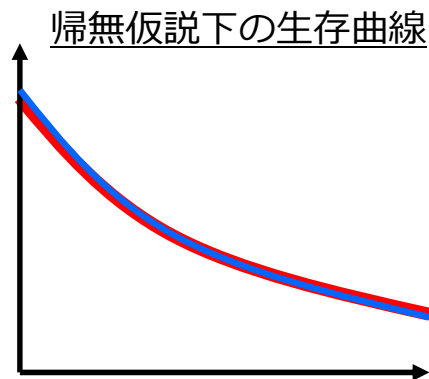
者はイベント発生時間に一定のずれが生じる場合に検出力が高くなることが知られている。しかしながら、がんペプチドワクチンにおいては、抗原特異的免疫応答を介した薬理薬効から遅発性の効果発現が想定されている。このことを考慮し、Harrington-Fleming 法²¹⁾で提唱されているような、観察期間後期に重み付けを置く、新規統計学的手法を用いた解析の必要性も想定される。統計学的解析方法については、有意と判断する規準とともにプロトコールに明記される必要がある。

日本バイオセラピー学会(2009). “がん治療用ペプチドワクチンガイダンス”: p.10.

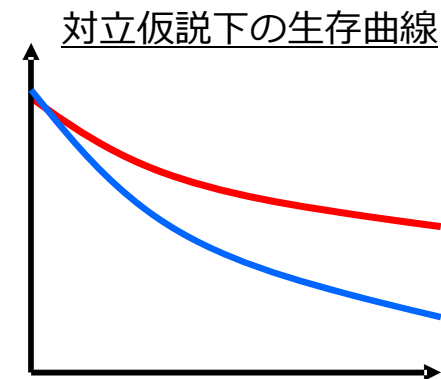
- 
- 比例ハザード性が成立しない場合(後述)
 - Cox回帰を用いたハザード比への要約は不適切
 - 「観察期間後期に重みをおいた検定」とは？
 - ログランク検定の拡張版検定
 - まず、ログランク検定の中身について理解しましょう！

[復習] ログランク検定

■ 生存曲線全体の比較



V.S.



- 帰無仮説の下で、
「**P値 = 試験結果以上の群間差が認められる確率**」を計算
- 事前に定めた基準(有意水準 α)と比較する

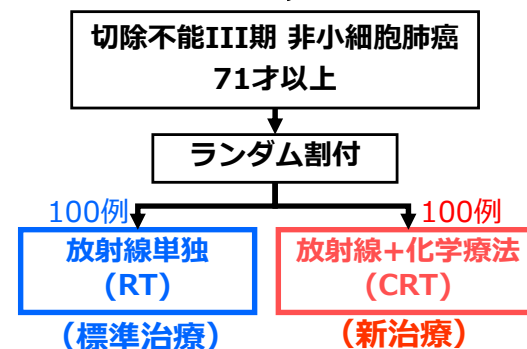
ログランク検定の中身

■ ある時点の死亡/打ち切りに関する2×2表

ある時点の# at risk

	死亡	打ち切り	計
RT	20	30	50
CRT	10	40	50

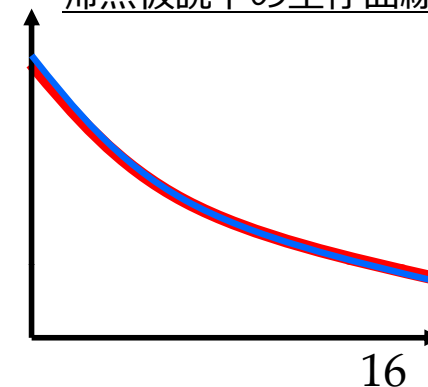
JCOG0301試験



■ 帰無仮説の下で想定される2×2表

	死亡	打ち切り	計
RT			50
CRT			50

帰無仮説下の生存曲線



■ 群を併合した場合の死亡割合

– $(10 + 20)/100 = 30\%$

■ 帰無仮説の下では両群の死亡割合は30%

- 両群とも死亡数は $50 \times 0.3 = 15$ 人と計算できる
- 帰無仮説の下で想定される2×2表

	死亡	打ち切り	計
RT	15	35	50
CRT	15	35	50

観測データ				帰無仮説下のデータ			
	死亡	打ち切り	計		死亡	打ち切り	計
RT	20	30	50	RT	15	35	50
CRT	10	40	50	CRT	15	35	50
観測イベント数				期待イベント数			

■観測イベント数と期待イベント数の差を計算

－ CRT群： $10 - 15 = -5$

■全イベント時点で計算して足しあわせる

－ “ログランクスコア”と呼ばれる

－ スコアが負の値をとった場合の解釈

- 全時点で平均的にCRT群の死亡数が少ない傾向
 - － CRT群の観測イベント数 < 期待イベント数
- 帰無仮説下ではスコアの期待値は0
 - － 0からの乖離が大きいほど死亡数が少ない傾向が強い

$$\text{ログランク検定統計量} = \frac{\text{スコアの 2乗}}{\text{スコアの分散}}$$

■ ログランク検定統計量

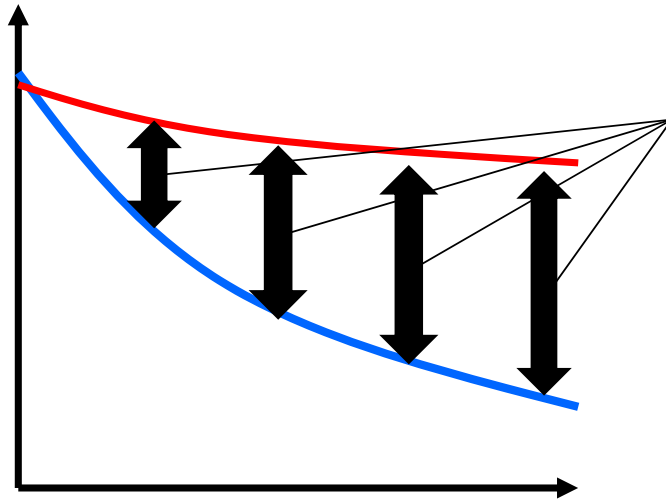
- 自由度1のカイ二乗分布にしたがう
 - カイ二乗分布からP値が計算可能

■ 検定の概略

- CRT群の死亡数はRT群より少ない @観測データ
 - 「少ない」 程度が大きくなると・・・
 - ⇒ ログランクスコアの2乗値が大きくなる
 - ⇒ 検定統計量も大きくなってP値が小さくなる

※ 詳細について興味のある方は、
「Green S, et al (JCOGデータセンター 訳). 米国SWOGに学がん臨床試験の実践－原著第3版」
の2.3節をご覧ください。

計算の過程から分かること



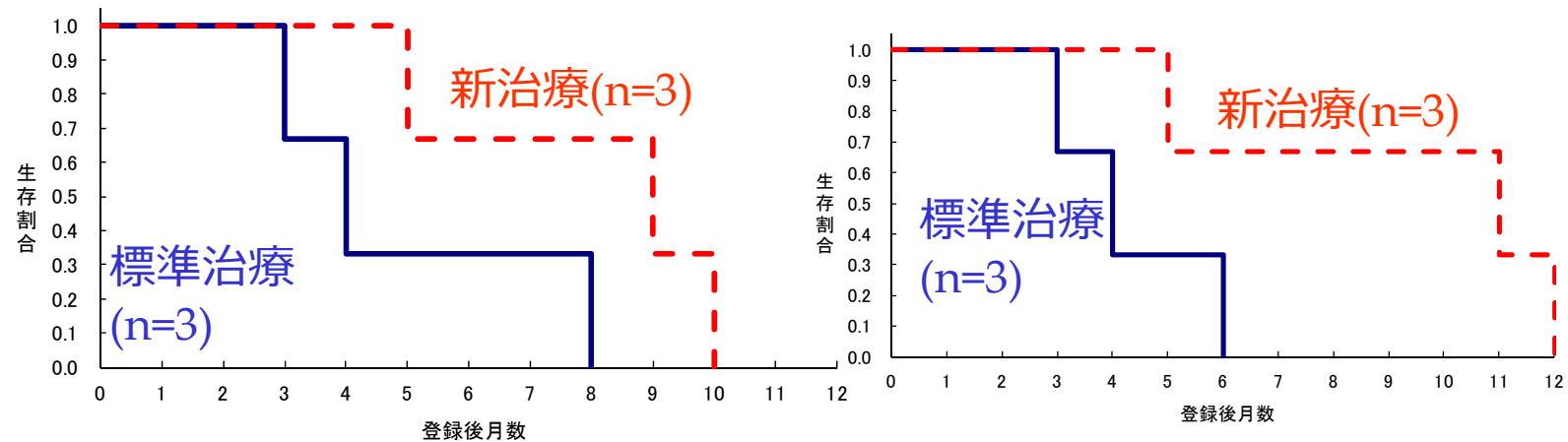
ロランク検定は・・・

- ・ イベント発生時点ごとの生存曲線のズレを足し併せて、曲線全体の差を比較している
- ・ “ズレ”を**全時点で等しく扱う**

ロランクスコア

$$\begin{aligned} &= (\text{第1イベント時点の観測イベント数} - \text{期待イベント数}) \\ &+ (\text{第2イベント時点の観測イベント数} - \text{期待イベント数}) \\ &+ (\text{第3イベント時点の観測イベント数} - \text{期待イベント数}) \\ &+ \dots \\ &+ (\text{最終イベント時点の観測イベント数} - \text{期待イベント数}) \end{aligned}$$

- イベントの発生順位の情報だけを用いる
- イベント発生までの時間の情報は用いない

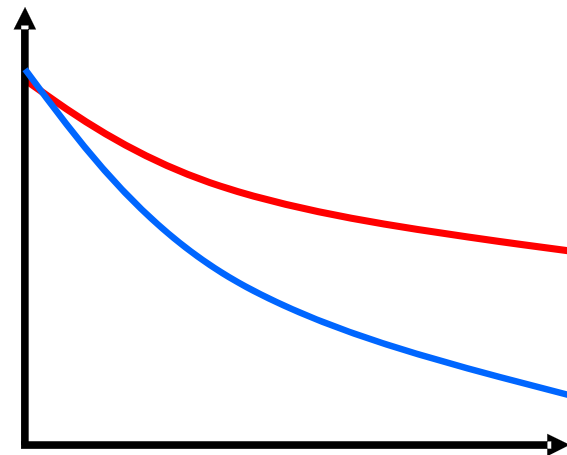


いずれも検定結果は同じ($P=0.108$)

ログランク検定の特徴

- 帰無仮説の下で誤って差があるとする確率(α エラー)が有意水準以下に保持される
 - “*delayed vaccine effect*”の有無によらず、帰無仮説において、ログランク検定を用いることは何ら問題はない
- 観測/期待イベント数の差を等しく扱う
 - **両群のハザードが全時点で比例関係にある場合に最適**
 - 真の生存曲線に比例ハザードタイプの差がある場合、検定結果が有意になりやすい(検出力が高い)
 - “*delayed vaccine effect*”が想定される場合、**ログランク検定を用いると検出力不足に陥る可能性がある**

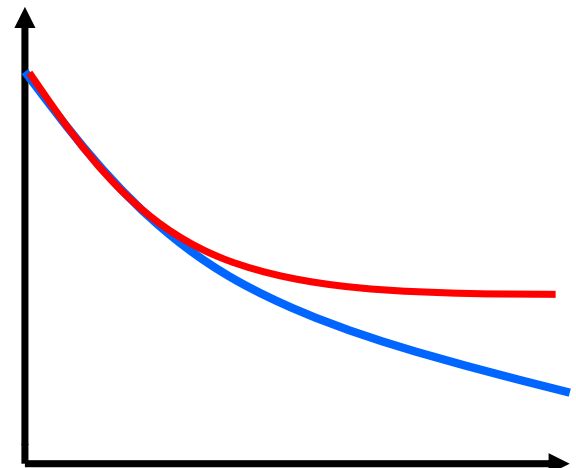
真の生存曲線の差(対立仮説での生存曲線)



比例ハザードタイプ



ログランク検定は有意になりやすい



delayed vaccine effect タイプ



ログランク検定は有意になりづらい
(後半の観測/期待イベント数の差に
大きな重みをおいた検定の方が
有意になりやすい)

※ 一般に、比例ハザードタイプでない試験は、
検出力不足に陥っている可能性があります
(免疫療法の臨床試験だけが特別な訳ではありません)

Fleming HR, Harrington DP. Counting Processes and Survival Analysis. John Wiley & Sons: New York, 1991

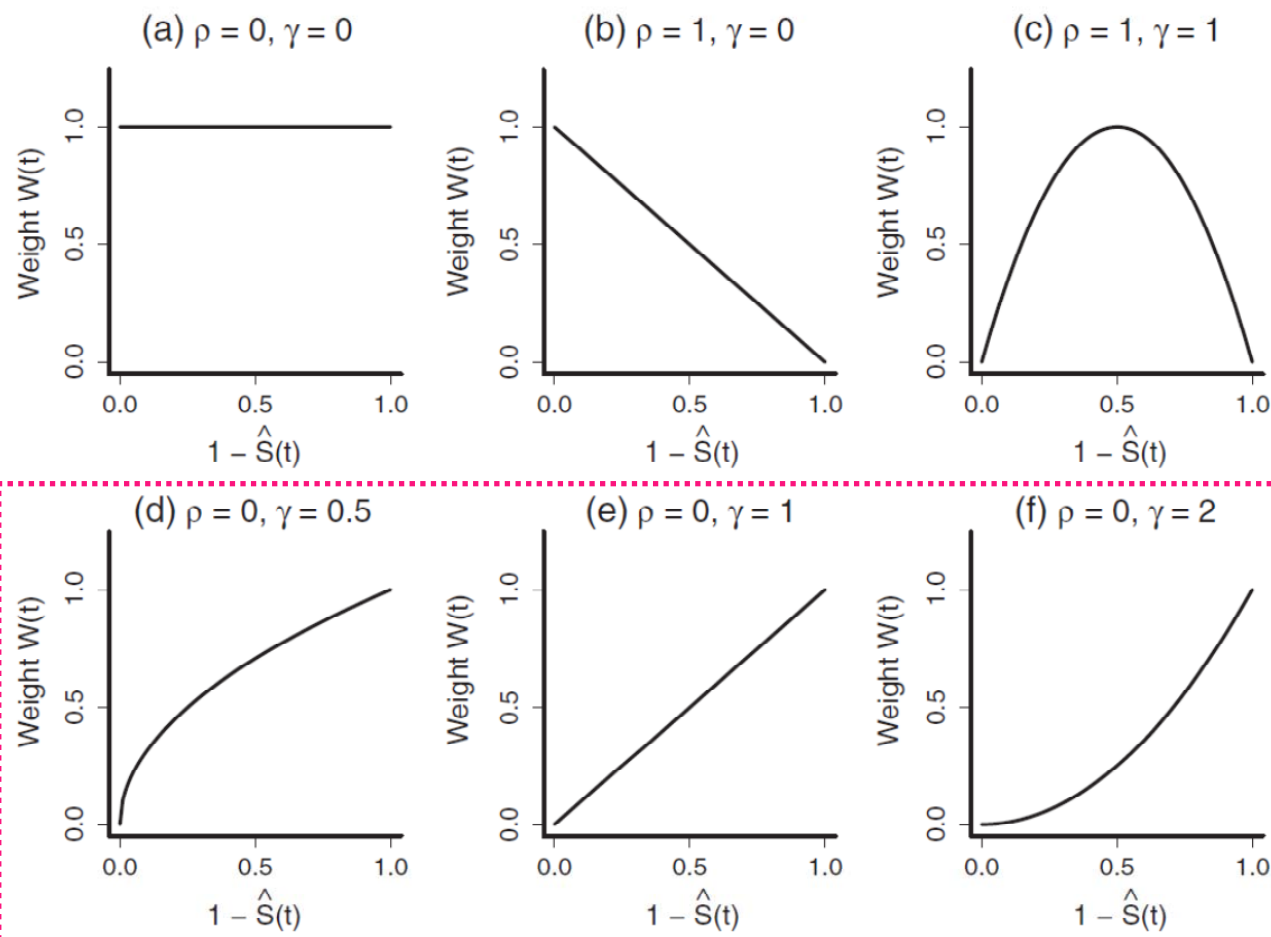
Fleming-Harrington検定(F-H検定)

- 一般化ロジランク検定の一種(G^p -family test)
 - 時点に応じて任意の重み付けを行った上でロジランクスコアを計算する

$$\text{F-H検定の重み} = S(t)^{\rho} \{1 - S(t)\}^{\gamma}$$

- $S(t)$: 時点 t の生存割合
- KM曲線の後半部分に大きな重みをおけば、
がん免疫療法タイプのKM曲線の検出力が向上する

F-H検定の重み



生存曲線の後半に大きな重みをおく場合

小括と補足

- “*delayed vaccine effect*”を検出したい試験
(主に薬効の証明(薬事承認)を意図した試験)
 - 検出力の観点では、F-H検定の方が適している
(**ログランク検定を使ってはいけない訳ではない**)

- 計画段階で確認/決定しておくべき事項
 - “*delayed vaccine effect*”が存在する証拠
 - F-H検定のパラメータ(ρ, γ)
 - パラメータに応じたサンプルサイズ設定

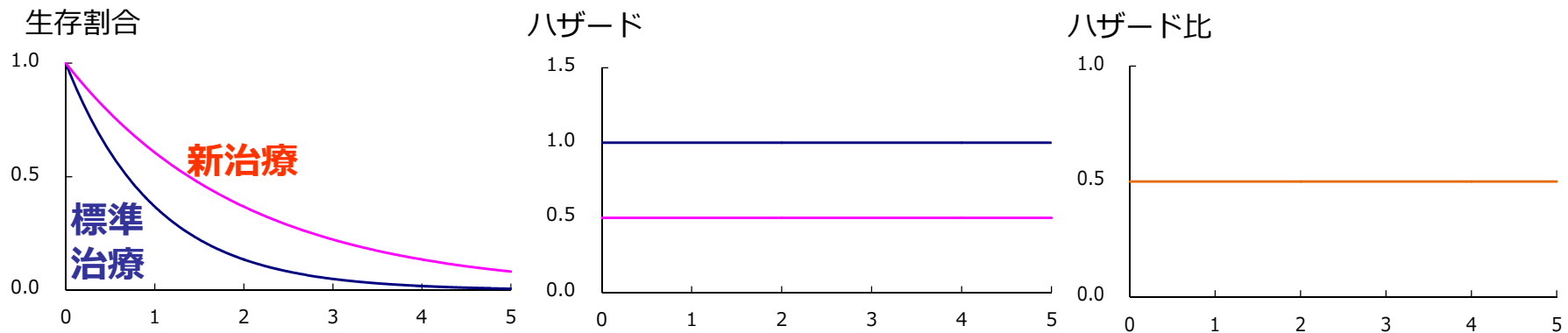
※ “*delayed vaccine effect*”が認められる時点やその前後における治療効果の大きさの違いなど、通常の試験より多くの仮定をおく必要があることに注意すべき

OUTLINE

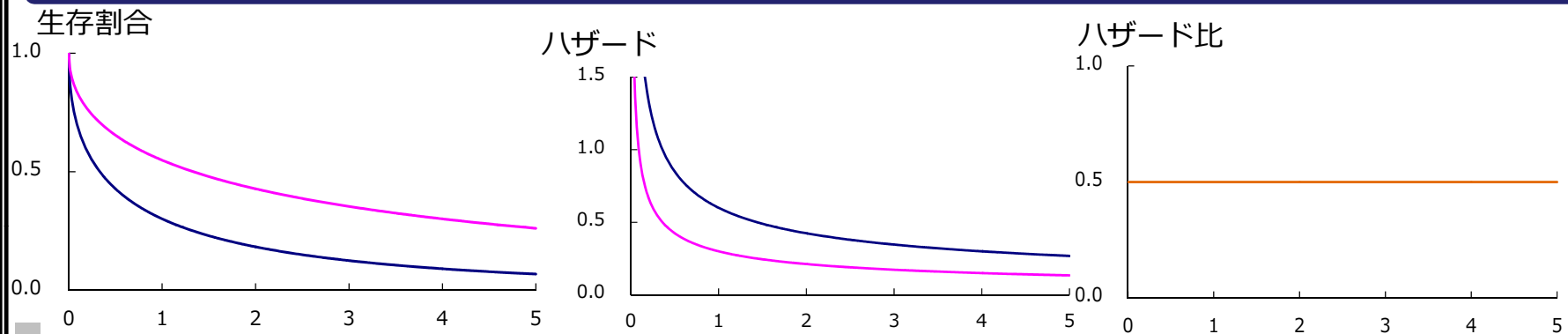
- Kaplan-Meier法は万能か？
 - ハザード関数
- ログラंक検定は万能か？
 - “delayed effect”の問題について
 - Fleming-Harrington検定の紹介
- Cox回帰は万能か？
 - ハザード比による要約は適切？

[復習] 比例ハザード性

指数分布：時点を通してハザードが一定。両群指数分布ならHRは一定



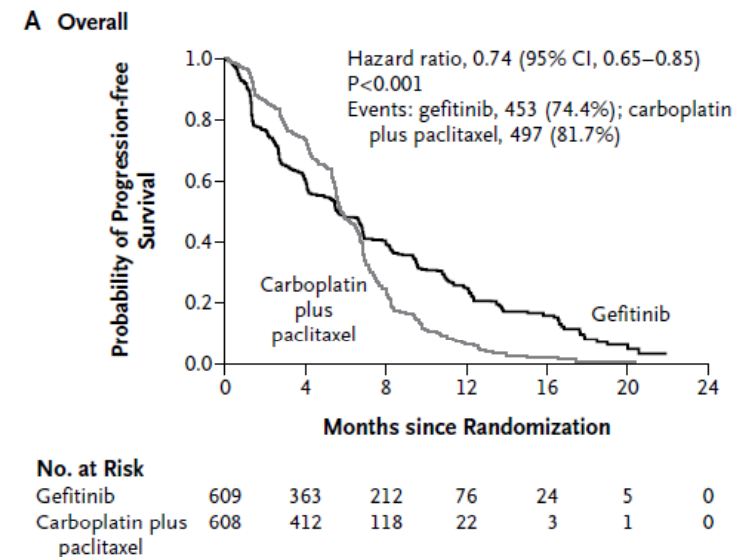
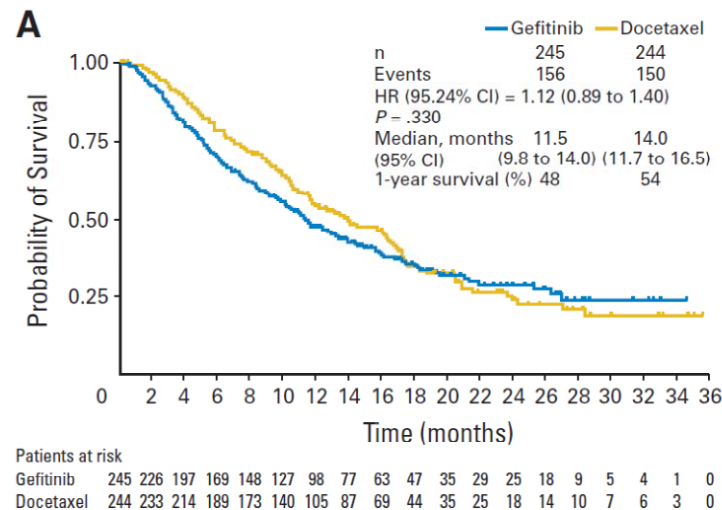
ハザードが一定でなくても、比例ハザード性は成立する(ワイブル分布の場合)



Maruyama R, et al. J Clin Oncol 2008. / Mok T, et al. N Engl J Med 2009.

ハザード比による要約が不適切な事例

■ NSCLCに対するV-15-32試験(左), IPASS試験(右)

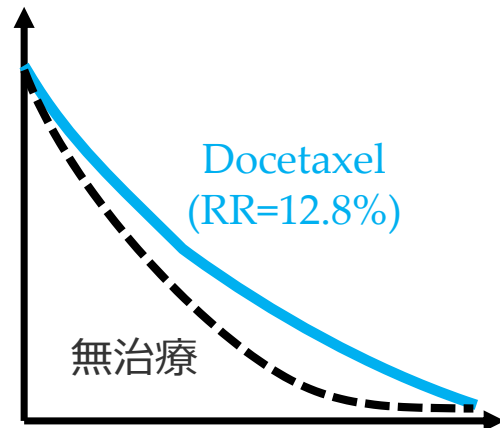


– OS/PFS曲線がクロスした理由とは？

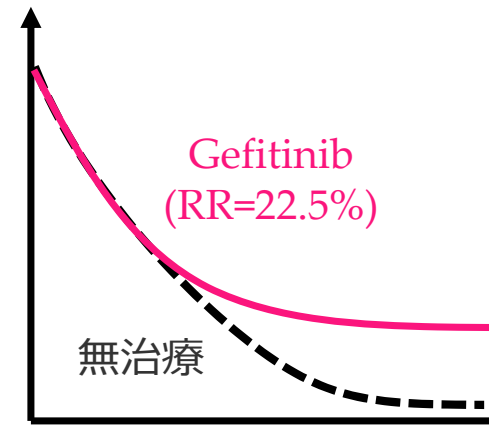
- EGFR変異と治療効果に(質的な)交互作用があった

里見清一(2011). “誰も教えてくれなかった癌臨床試験の正しい解釈”: p.93-96.

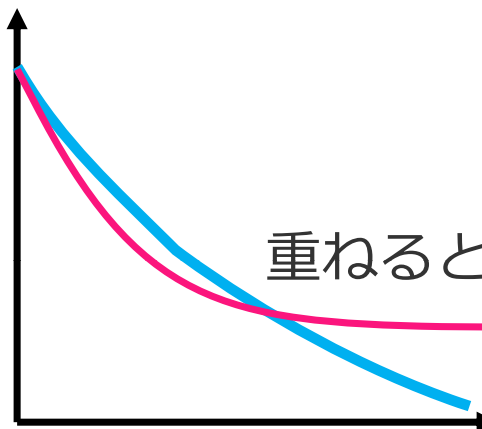
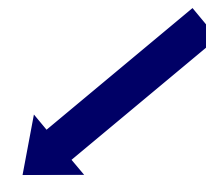
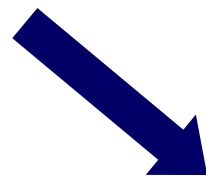
V-15-32試験結果の考察



- 全ての患者にちよつとずつ効く
- 長期生存例はいない



- 一部の患者にかなり効く
- 長期生存例が存在する

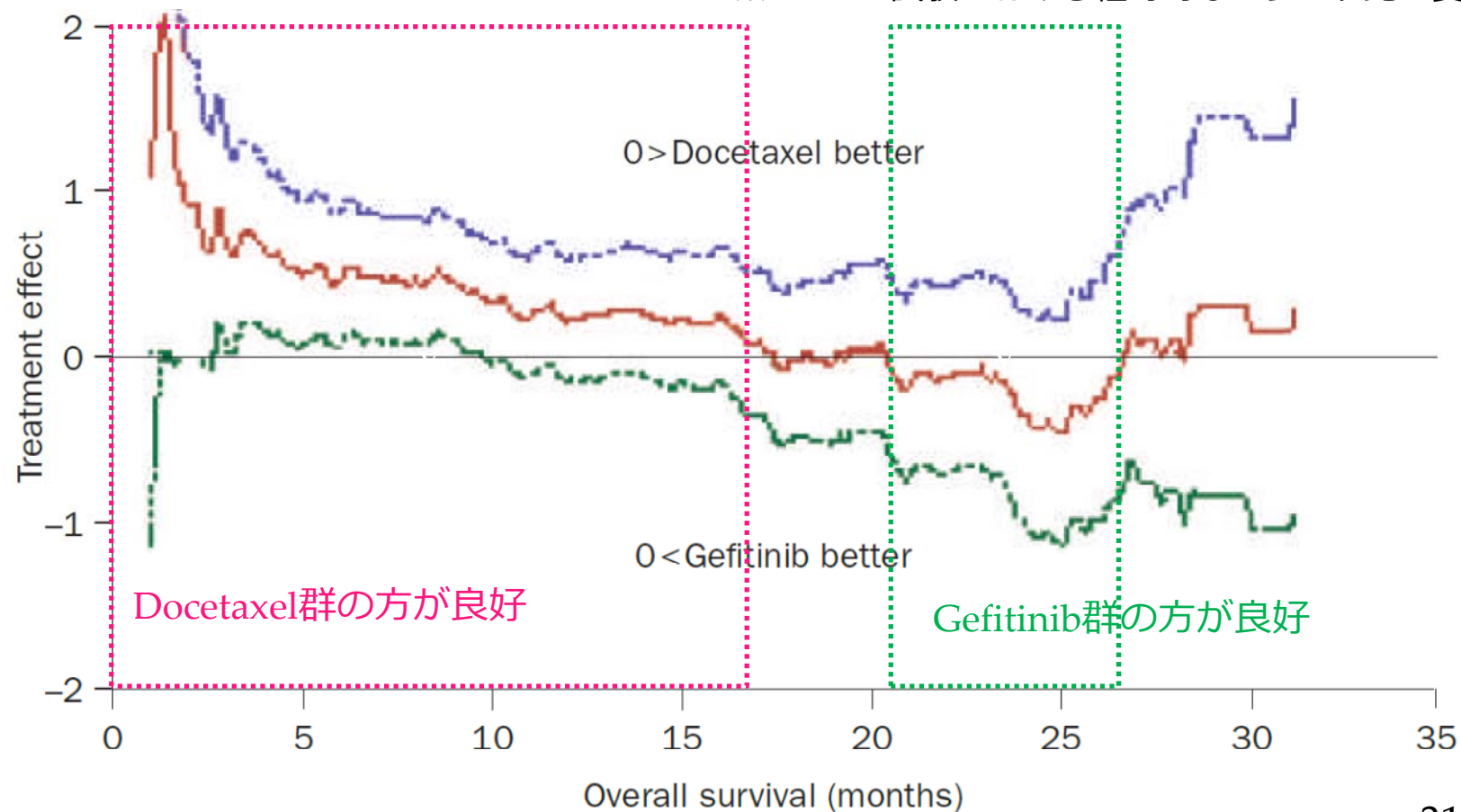


重ねると生存曲線がクロス！

Saijo N, et al. Nat Rev Clin Oncol 2009.のFig3を改変

交互作用が存在する場合の(対数)ハザード比

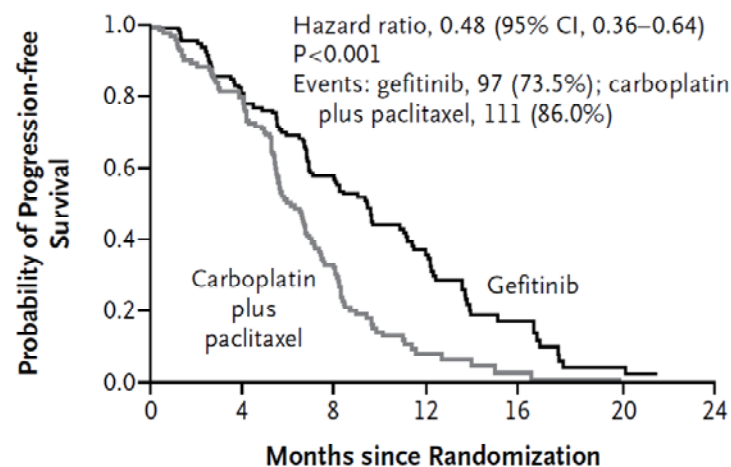
※ V-15-32試験における経時的なハザード比の変化



Mok T, et al. N Engl J Med 2009.のFig2より抜粋

IPASS試験における交互作用

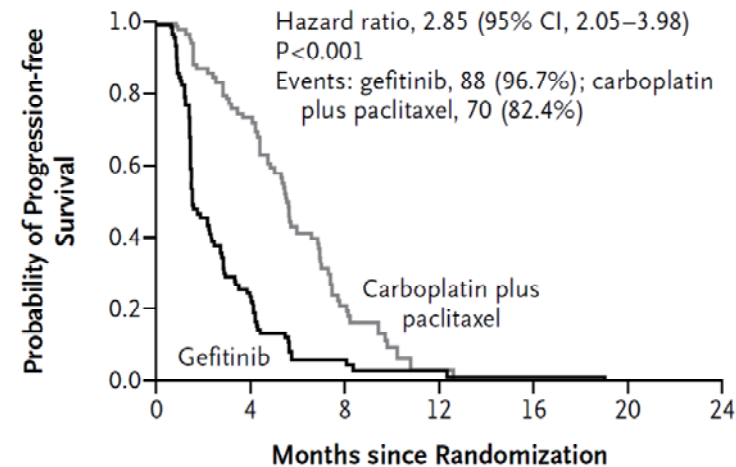
B EGFR-Mutation-Positive



No. at Risk

Gefitinib	132	108	71	31	11	3	0
Carboplatin plus paclitaxel	129	103	37	7	2	1	0

C EGFR-Mutation-Negative



No. at Risk

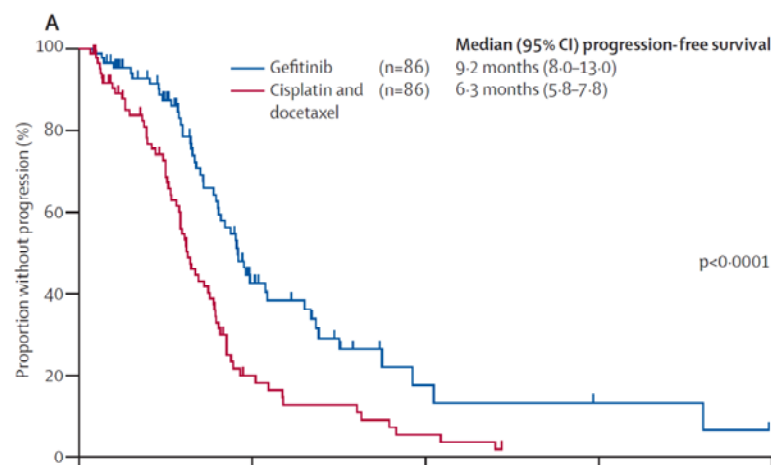
Gefitinib	91	21	4	2	1	0	0
Carboplatin plus paclitaxel	85	58	14	1	0	0	0

質的な交互作用が存在する場合はKM曲線はクロスする

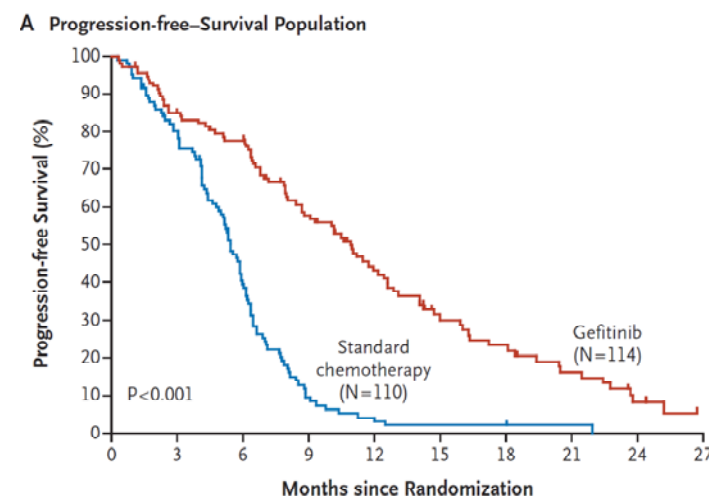
有効なサブセットが存在する場合

■ マーカーを考慮したデザインを採用すべき

— 対象を絞った場合、質的な交互作用は認められなかった



Mitsudomi T, et al. Lancet Oncol 2010.



Maemondo M, et al. N Engl J Med 2010.

— 種々のデザイン・解析アプローチが考案されている

里見清一(2011). “誰も教えてくれなかった癌臨床試験の正しい解釈”: p.111-129.

Simon R(2013). “Genomic Clinical Trials and Predictive Medicine”.

Freidlin, B., Sun, Z., Gray, R., & Korn, E. L. (2013). Phase III Clinical Trials That Integrate Treatment and Biomarker Evaluation. *Journal of Clinical Oncology*, 31(25), 3158-3161.

Uno H, et al. J Clin Oncol 2014.

ハザード比に代わる要約指標の提案

VOLUME 32 • NUMBER 22 • AUGUST 1 2014

JOURNAL OF CLINICAL ONCOLOGY

STATISTICS IN ONCOLOGY

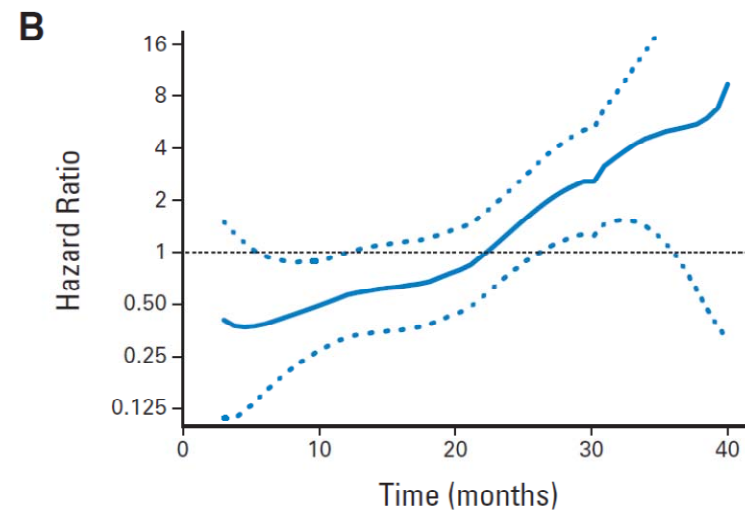
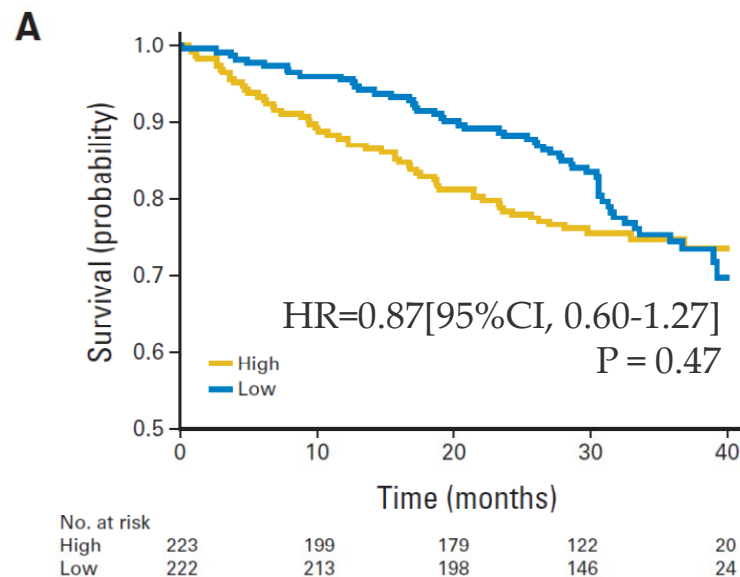
Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis

*Hajime Uno, Brian Claggett, Lu Tian, Eisuke Inoue, Paul Gallo, Toshio Miyata, Deborah Schrag,
Masahiro Takeuchi, Yoshiaki Uyama, Lihui Zhao, Hicham Skali, Scott Solomon, Susanna Jacobus,
Michael Hughes, Milton Packer, and Lee-Jen Wei*

- 比例ハザード性に問題がある場合を想定し、
ハザード比以外の要約指標を提案

Uno H, et al. J Clin Oncol 2014.のFig1を改変

- multiple myelomaに対するdexamethasoneの試験
– ハザード比とOSの間に質的な交互作用を認めた(右図)
• Low dose群は試験前半で予後良好



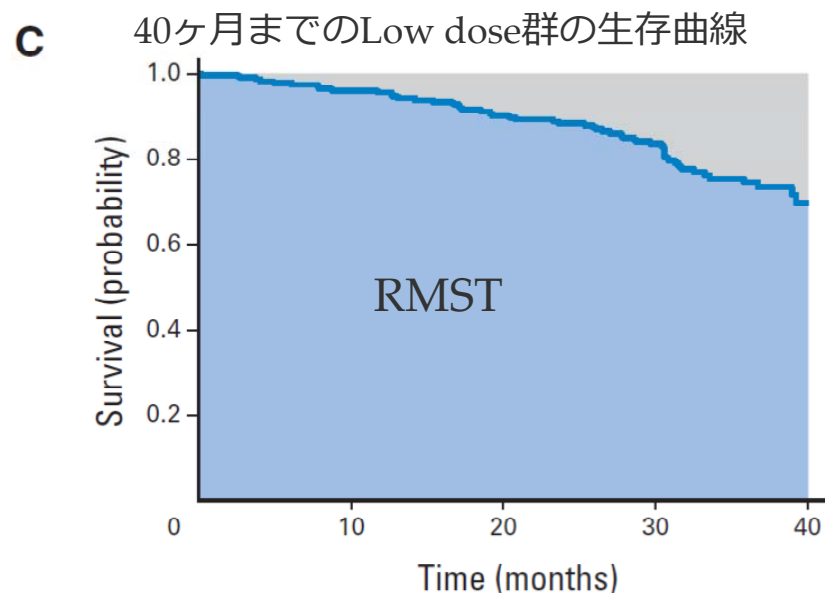
Uno H, et al. J Clin Oncol 2014. ■■

Uno H, et alの提案

- X年生存割合の比または差
 - 臨床的に関心のあるX年で評価する
- KM曲線の分位点(50%点ならMST)
- **Restricted mean survival time (RMST)**の比または差
 - KM曲線のX年までの曲線下面積
 - X年までの生存期間の算術平均に一致
 - 曲線より上の面積(restricted mean time lost)で評価することも提案

Uno H, et al. J Clin Oncol 2014.のFig1/Table1より改変

HRと提案指標の比較



Group Contrast Measure	Rajkumar et al ¹⁰ (myeloma)	
	Estimate	95% CI
Hazard ratio (PH model)	0.87	0.60 to 1.27
<i>t</i> -year survival	Month 40	
Difference	-0.04	-0.15 to 0.06
Ratio	0.95	0.82 to 1.10
Percentiles	10th	
Difference (months)	10.9	2.6 to 19.1
Ratio	2.15	1.17 to 3.96
Restricted mean survival time	Month 40	
Difference (months)	2.2	0.1 to 4.2
Ratio	1.06	1.00 to 1.13
Ratio of restricted mean time lost	0.68	0.47 to 0.98

Abbreviations: NSCLC, non-small-cell lung cancer; PH, proportional hazards.

RMSTでは差・比ともにLow dose群が上回っている

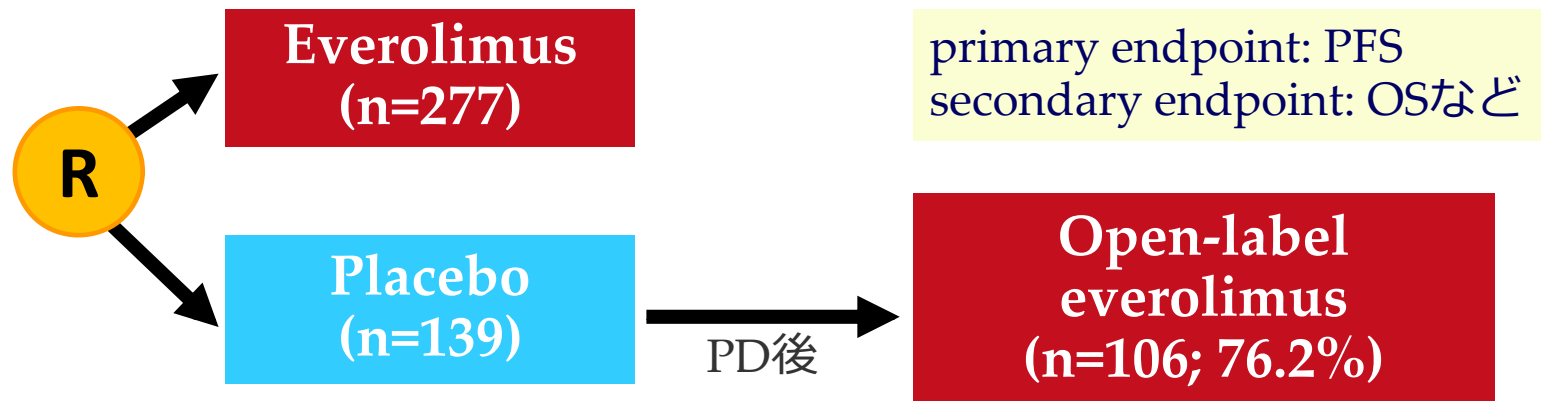
有効な後治療が存在するケース

■ 腎細胞癌に対するRECORD-1試験

- 早期試験でmTOR阻害剤の高い有効性を確認

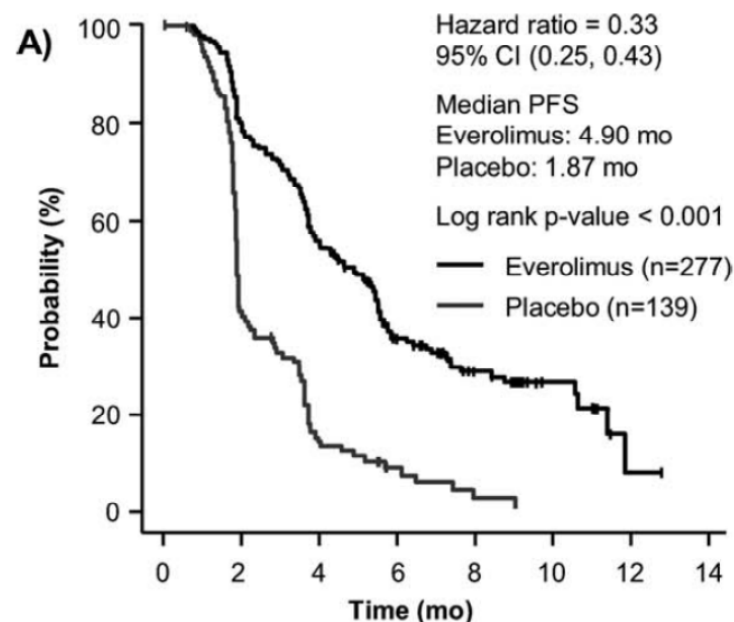
(Amato RJ, et al. ASCO 2006. abst # 4530)

- placebo群でPDを認めた患者はeverolimusのopen label試験への登録が認められた



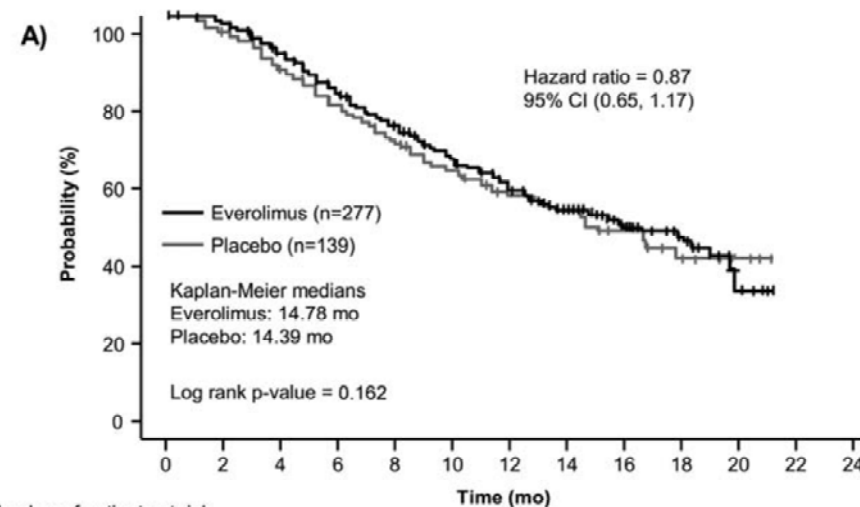
Motzer RJ, et al. Cancer 2010. PFS(Fig 2A), OS(Fig 6A)

RECORD-1試験のPFS(左図)とOS(右図)



Number of patients at risk

Everolimus	277	192	115	51	26	10	1	0
Placebo	139	47	15	6	2	0	0	0



Number of patients at risk

Everolimus	277	267	240	204	164	155	131	101	61	30	6	0	0
Placebo	139	131	117	100	88	74	56	43	27	13	3	0	0

everolimusに延命効果はないのか？
76%ものクロスオーバーの影響を受けてしまったからでは？

クロスオーバーを補正する方法

■ バイアスが入ってしまう(が頻用される)手法

- クロスオーバー例を除外あるいは打ち切り

■ 因果推論に基づく手法

- “クロスオーバーがなかったら・・・”
という仮定の下で治療効果を推定する方法

- **Inverse probability of censoring weighted method (IPCW法)**

(Rimawi M and Hilsenbeck SG. J Clin Oncol 2012.)

- **Rank-preserving structural failure time (RPSFT) model**

(Korn E and Freidlin B. J Clin Oncol 2011.

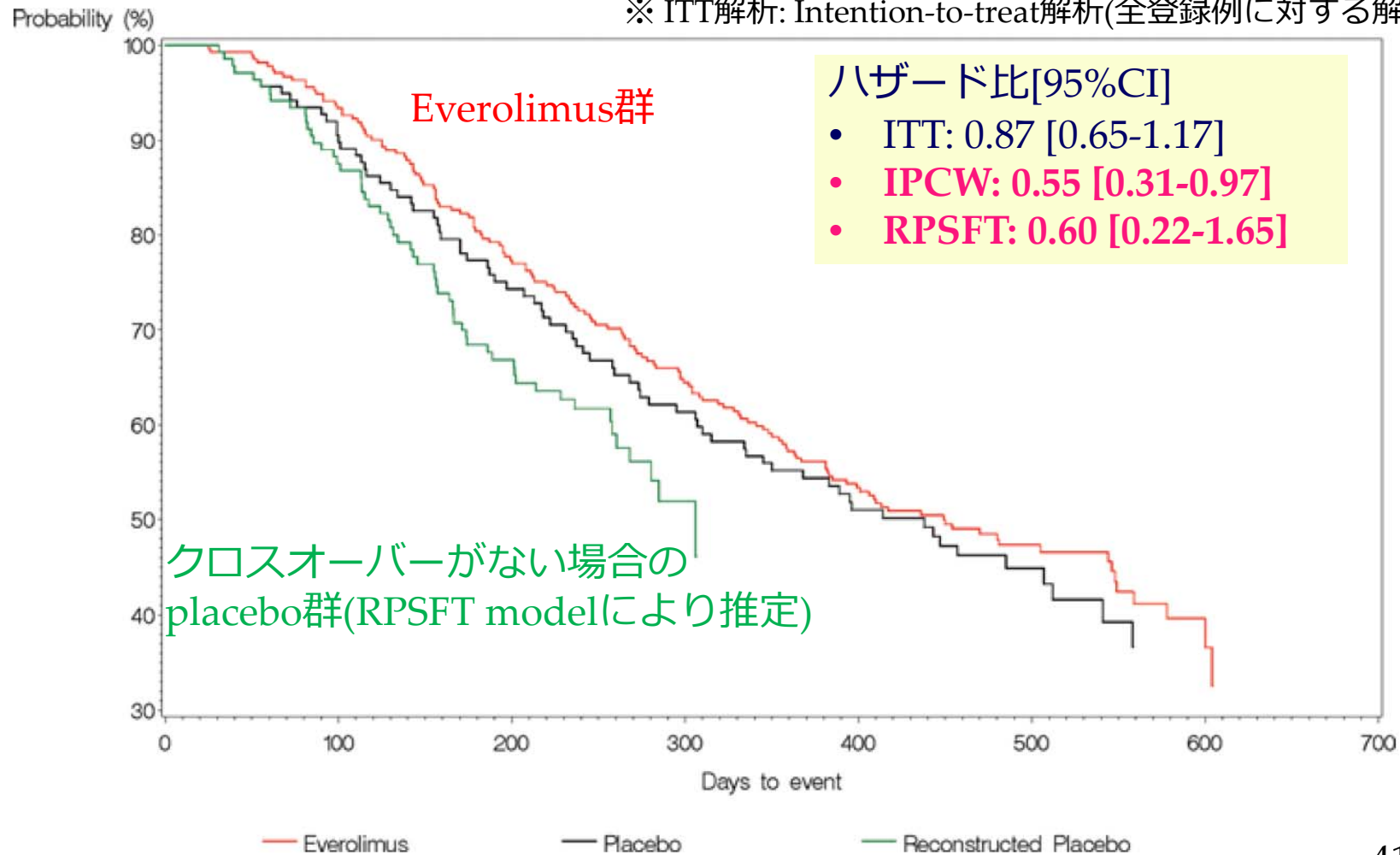
/ Freidlin B and Korn E. Bone Marrow Transplant 2011.)

※ 因果推論に基づく手法は多くの仮定が必要であることに注意！

Korhonen P. J Biopharm Stat 2012. / NICE (2011). Technology appraisal guidance 219.

RECORD-1試験を因果推論の手法で解析した結果(OS)

※ ITT解析: Intention-to-treat解析(全登録例に対する解析)



Korn E, et al. J Clin Oncol 2011. / Buyse M, et al. J Clin Oncol 2011. / Korn E, et al. J Clin Oncol 2011.

有効な後治療が存在する下でのOSの評価

※ クロスオーバーする試験治療を「後治療」と呼んでいます

■ 後治療が既にpracticeとして選択可能な場合

－ ハザード比による要約は適切

- pragmaticな比較に興味がある：「先にやるか後にやるか？」
- クロスオーバーを補正する必要はない

■ 後治療がpracticeで選択可能でない場合

－ 科学的にはクロスオーバーを禁止したいが、 倫理的に許容できないケースもある(特に新薬のpivotal試験)

- クロスオーバーの影響を受けないendpoint (例えばPFS)や、
因果推論に基づく手法の利用が有用

ハザード比による要約は適切？

■ 比例ハザード性に問題がある場合

- 生存曲線がクロスしているケース
 - 交互作用の検出を意図したサブグループ解析が有用
 - 事前に分かるならマーカーを加味したデザインを利用するのも一案
- 治療効果の高いサブセットがないケースでは、RMSTなどの指標を用いることも一案

※ ITT: Intention-to-treat解析(全登録例に対する解析)

■ practiceで選択できない試験治療へのクロスオーバーが避けられず、その頻度が高い場合

- ITT解析のハザード比からは治療効果が分からない
- ハザード比の信頼区間が広くなり、検出力が低下
- 因果推論の手法を用いて検討することは一案

まとめ

■ “3種の神器”は万能でない

－ ハザード関数を紹介

- リスクの経時的変化が把握できる
- 指数分布や比例ハザード性の仮定について検討できる

－ F-H検定を紹介

- 後半の差を見いだすことに意義・妥当性があるなら有用
- 多くの場合、仮定の少ないロジランク検定が選択されるべき

－ ハザード比による要約が不適切なケース

- 比例ハザード性に問題がある場合
 - － 治療効果の高いサブセットが存在する場合 → マーカーを考慮したデザイン
 - － 有効なサブセットが存在しない場合 → RMSTを紹介
- 有効な試験治療へのクロスオーバーが高頻度に発生する場合
 - － 因果推論の手法やPFSの情報が有用となりうる

Further reading

■ ハザード関数について

Hess, K. R., & Levin, V. A. (2014). Getting More Out of Survival Data by Using the Hazard Function. *Clinical Cancer Research*, 20(6), 1404-1409.

■ gefitinibの試験の考察

Saijo, N., Takeuchi, M., & Kunitoh, H. (2009). Reasons for Response Differences Seen in the V15-32, INTEREST and IPASS Trials. *Nature Reviews Clinical Oncology*, 6(5), 287-294.

里見清一(2011). “誰も教えてくれなかった癌臨床試験の正しい解釈”: p.93-96.

大橋靖雄(2011). “医師のための臨床統計学 基礎編”: p.118-122.

■ ハザード比の代わりとなる指標・解析法

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., ... & Wei, L. J. (2014). Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis. *Journal of Clinical Oncology*, 32(22), 2380-2385.

Freidlin, B., & Korn, E. L. (2011). Assessing Causal Relationships Between Treatments and Clinical Outcomes: Always Read the Fine Print. *Bone Marrow Transplantation*, 47(5), 626-632.