癌臨床試験における無作為化比較第II相試験のデザインについて

下川 敏雄

和歌山県立医科大学 臨床研究センター

本レクチャーの流れ

- 無作為化比較第II相試験の動機
- 無作為化比較第II相試験のデザイン
  - 並行非比較レジメン
  - ランダム化選択デザイン
  - ランダム化スクリーニングデザイン
  - ランダム化中止デザイン
- 単アーム試験 vs. 無作為化比較試験

おはなしの前に：がん臨床試験のフェーズ

Phase I/II

Phase I

用量反応毒性の検討

最大耐用量(MTD)がわかる

Single-arm phase II

Historical Controlとの比較

非比較の範疇で治療効果(effect size)がわかる

Randomized phase II

複数の試験治療アームの比較

Selection biasを除外したもとで治療効果がわかる(かも)

Seamless phase II/III

Phase III

標準治療との比較

標準治療に対する有効性・安全性が検証できる

なぜ無作為化比較第II相試験が必要か？

単アーム試験(Historical controlとの比較)の限界

- 分子標的薬の場合には，標準治療がCytotoxic drugでの結果であるときに，HER2, EGFR,VGFR等のサブセットの情報がないため，対象母集団が異なる可能性がある(標準治療の結果に疑義がもたれる).
- ヒストリカルコントロールの試験時に比べて，診断技術や他の医療技術の発展が進み，当該治療以外の要因が含まれる可能性がある.
- そもそもヒストリカルコントロールが存在しない(研究対象下では存在しない場合)には単アーム試験の計画が困難である(例えば，院内の臨床データを利用した場合には，多施設共同試験で実施するときには施設バイアスを伴ってしまう).
- とくに，全生存期間(OS)の場合にはその傾向は顕著であり，また，少数例での結果であるため，ばらつきが大きく信頼性が低い可能性がある.

無作為化比較第II相試験を実施することで，上記問題は解決できる可能性があり，大規模な第III相試験での試験成績の向上に繋がるかもしれない(第III相試験は難しいので第II相というのではダメ)．さらに，効果予測因子の探索にもつながる可能性がある.

単アーム第II相試験に対する批判

Zia(2005)は181件のPIII試験を調査し，そのなかで同一レジメンで同一集団に対してPII試験が実施された43試験に関して調査を行った.

Zia et al.: Journal of Clinical Oncology, 28, 6982-6990, 2005.

		Phase III (Total=43)	Phase II (Total=49 [RPII = 2])
癌種	肺癌	15 (35%)	
	乳癌	9 (21%)	
	大腸癌	7 (16%)	
	悪性黒色腫	5 (12%)	
	その他	7 (16%)	
結果	Positive	12 (28%)	
	Negative	31 (72%)	
症例数	平均[SD]	363 [209]	52 [31.9]
	範囲	154 - 1,155	12 - 137
RR	平均[SD]	**34.2 [19.7]**	**43.2 [19.8]**
	範囲	**10.8 - 85.7**	**16.0 - 87.0**

直線の下に●(臨床試験を表す)が集中している(35/43試験)．すなわち，PIII試験では，PII試験に比べて，RRが大幅に落ちていることがわかる.

- 標本サイズが小さい(結果の(統計学的な)精度)
- 検定が保守的でない(α，βエラーの設定の問題)
- 参加施設拡大に伴う試験計画の遵守，プロトコル治療に関する慣れ(施設バイアス)
- (生存期間の場合)フォローアップ期間等の短さ

さらに，分子標的薬が主流となった現在，HER2, EGFR, VGFRサブセットでの情報が必要になっているものの，それらがHistorical Controlではとられておらず，全体としてのデータを用いることがある(対象母集団の違い).

無作為化比較第II相試験が注目されるきっかけになった試験

膵癌に対する1st-line治療としてのgemcitabine (GEM)をヒストリカルコントロールとした第II相試験において，いくつかのレジメンに対するpositive studyが報告された．そのため，gemcitabineを対照とした第III相試験が実施された(Green et al., 2012[1])．

vs. GEM+Ceuximab (Philip et al., 2010[2])

Xiong et al.(2004) では，41例のEGFR(+)の膵癌患者に対して，GEM+Cetuximabを実施した．その結果，中央生存期間(MST)は7.1カ月であった．

HR (GEM/GEM+α)=1.33 (OS)

GEM alone (MST=6M),GEM+Cetuximab (MST=8M)

Xiong et al.: Journal of Clinical Oncology, 22, 2610-2616, 2004.

vs. GEM+Bevacizumab (Kindler et al., 2010[3])

Kindler et al.(2005) では，52例のVEGR(+)の膵癌患者に対して，GEM+Bevacizumabを実施した．その結果，中央生存期間(MST)は8.8カ月であった．

HR (GEM/GEM+α)=1.38 (OS)

GEM alone (MST=6M),GEML+Bevacizumab (MST=8.8M)

Kindler et al.: Journal of Clinical Oncology, 24, 8033-8040, 2005.

vs. GEM+Oxaliplatin (Louvet et al., 2005[4])

Louvet et al.(2002) では，62例の進行膵癌患者に対して，GEM+Oxaliplatinを実施した．その結果，中央生存期間(MST)は9.2カ月であった．

8カ月全生存率

GEM alone = 30%, GEMOX = 50%

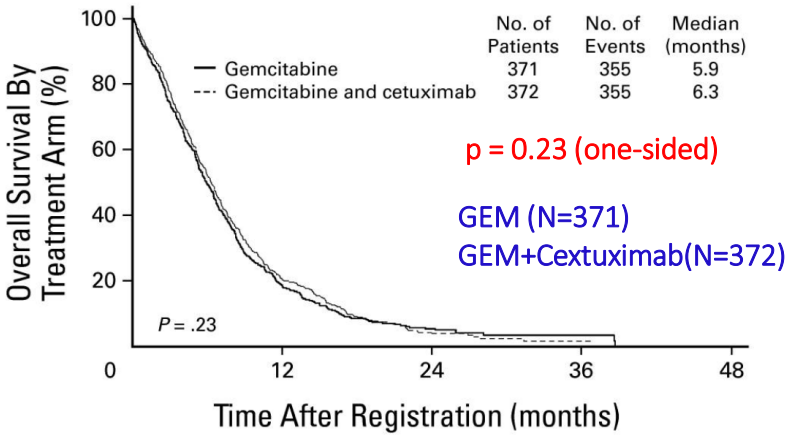
Louvet et al.: Journal of Clinical Oncology, 22, 1512-1518, 2002.

1: Green et al.: Clinical Trials in Oncology (3rd Edition), CRC Press, 2012.
2: Philip et al.: Journal of Clinical Oncology, 28, 3605-3610, 2010.
3: Kindler et al.: Journal of Clinical Oncology, 28, 3617-3622, 2010.
4: Louvet et al.: Journal of Clinical Oncology, 3509-3516, 2005.

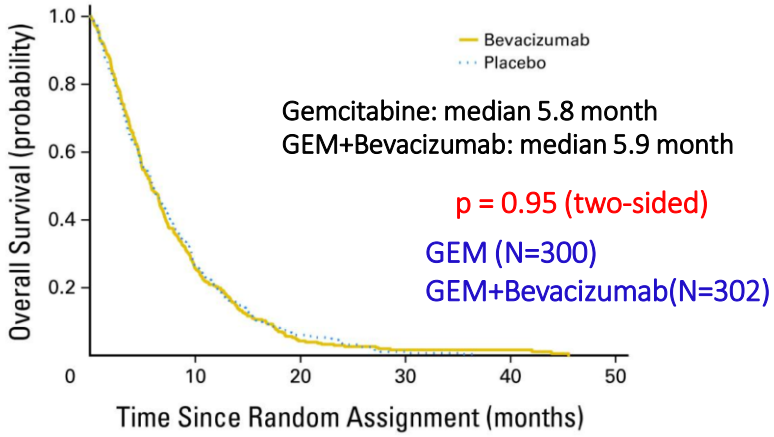
無作為化比較第II相試験が注目されるきっかけになった試験

膵癌に対する1st-line治療としてのgemcitabine (GEM)をヒストリカルコントロールとした第II相試験において，いくつかのレジメンに対するpositive studyが報告された．そのため，gemcitabineを対照とした第III相試験が実施された(Green et al., 2012[1])．

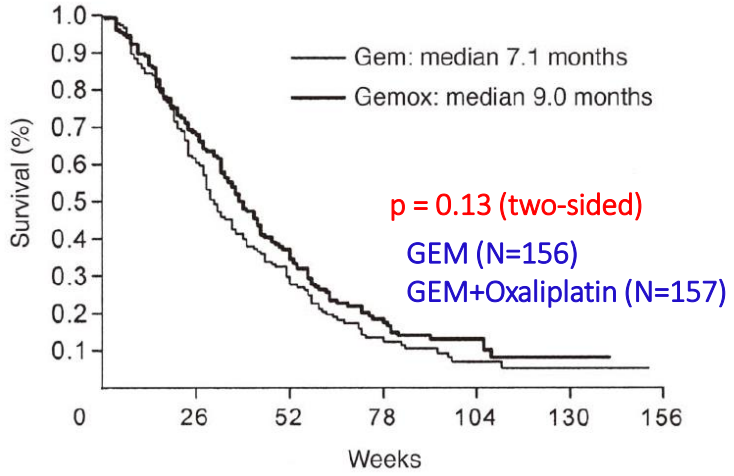
vs. GEM+Cextuximab (Philip et al., 2010[2])



vs. GEM+Bevacizumab (Kindler et al., 2010[3])



vs. GEM+Oxaliplatin (Louvet et al., 2005[4])



いずれの試験もポジティブスタディとはならなかった．すなわち，単アーム試験での結果をもって第III相試験を実施することに対することに疑義が生じた．

1: Green et al.: Clinical Trials in Oncology (3rd Edition), CRC Press, 2012.
2: Philip et al.: Journal of Clinical Oncology, 28, 3605-3610, 2010.
3: Kindler et al.: Journal of Clinical Oncology, 28, 3617-3622, 2010.
4: Louvet et al.: Journal of Clinical Oncology, 3509-3516, 2005.

第II相試験においてPositiveなときに

Positiveな第II相試験に続いて行われる第III相試験での偽陽性確率[1)]

Crowley and Hotering (2012)は，第II相試験において実施される治療レジメンのなかで，真に有効である確率を考えたうえで，偽陽率(有効だと判断しにも関わらず，本当は有効でない確率)をBayesの定理で表している．

第1種の過誤α 第2種の過誤β	試験治療が有効なときに有効だと判断する確率(1-β)	治療レジメンが真に有効である確率が10%であると仮定したもとでの偽陽率	治療レジメンが真に有効である確率が20%であると仮定したもとでの偽陽率
0.05, 0.10	90%	33%	18%
0.10, 0.10	90%	50%	31%
0.10, 0.20	80%	53%	33%
0.20, 0.20	80%	69%	50%

もし，治療レジメンが有効である確率(試験治療の有効確率)が10%であると仮定するとき，有意水準α=0.05, 検出力1-β=0.90のもとで試験を実施しても，本来は有効な治療アーム(Active)なのに「有効でない」と誤る(有効な治療を世に埋もれさせてしまう)確率は33%もある．

「有効である」との判断を第III相試験と考えると，いかに第II相試験から第III相試験に移行した後に有効性を示すことが難しいかを見ることができる．

1：Crowley, J. and Hotering, A. (eds.): Handbook of Statistics in Clinical Oncology (3rd Edition), CRC Press, 2012.

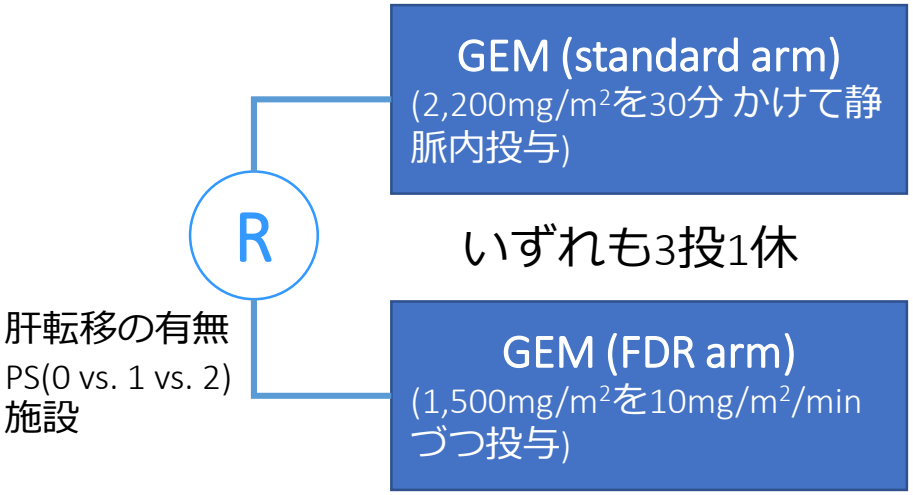
無作為化比較第II相試験に対する批判

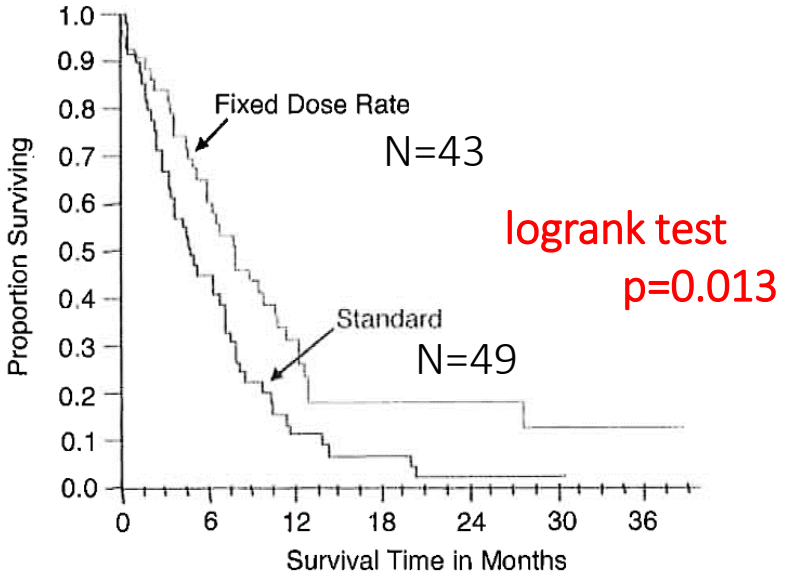
膵癌に対してgemcitabineを通常より暖徐に点滴して一定以上の血中濃度を長時間保つという方法を通常のgemcitabine投与法と比較したrandomized phase II trialがある．・・・2群あわせて90例ちょっとという少ない例数にも関わらず，OSは有意差をもって(p=0.013)長期間点滴のほうが良好であった．しかしこの長期間点滴法については，その後の800例以上のphase III trialで効果が否定されている．

やはりrandomized phase IIでの「有意差」はあてにならない(里見, 2011).

この書籍において批判された無作為化比較第II相試験(Tempero et al., 2003)

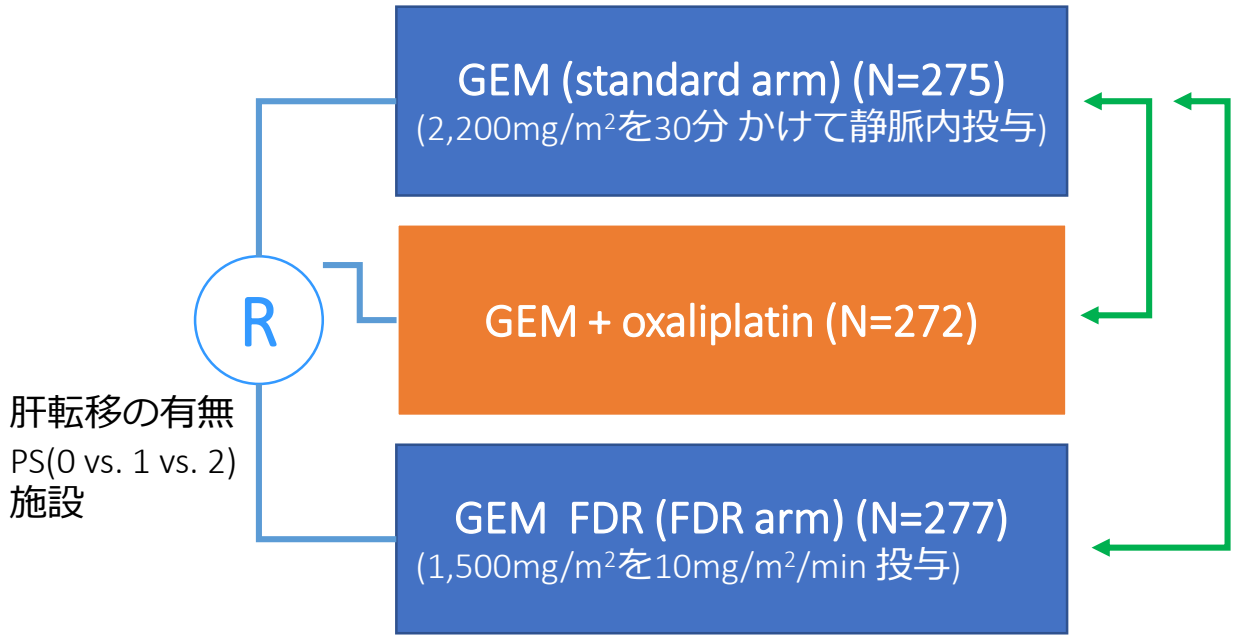
Tempero et al. :Journal of Clinical Oncology, 3402-3408, 2003.



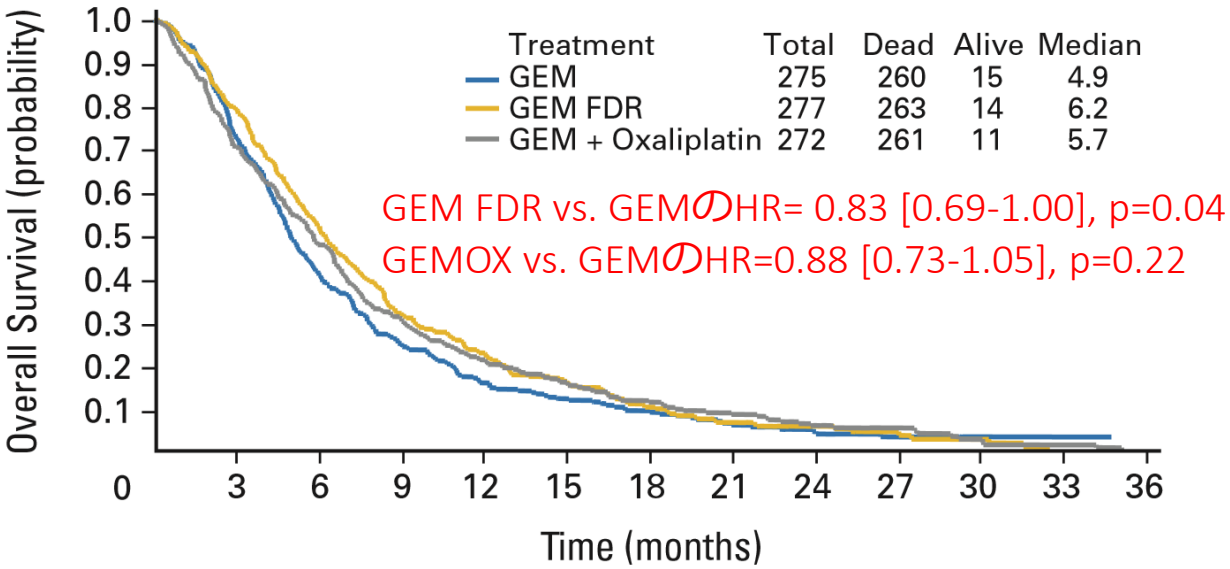


ただし，Poplin et al. (2009)では，negative studyだった.

Poplin et al. (2009)におけるスタディデザインと結果



MSTにおいて33%の差が得られるように設定．αエラー(両側)を全体で5%とするために，1比較当たり2.5%(多重比較)となる(検出力1-βは80%)．



有意水準α=0.025なので，有意ではない．ただし，2armでの試験だったとするならば，GEM vs. GEM FDRがpositive studyだったかもしれない．したがって，「まったくあてにならないわけでない」

無作為化比較第II相試験のデザイン

Mandrekar and Sargent(2010)[1)]は，無作為化比較第II相試験のタイプとして，以下を提示している.

- **並行非比較レジメン** (Parallel Noncomparative Regimens)
- **ランダム化選択デザイン** (Randomized Selection Designs)
- **ランダム化スクリーニングデザイン** (Randomized screening Designs)

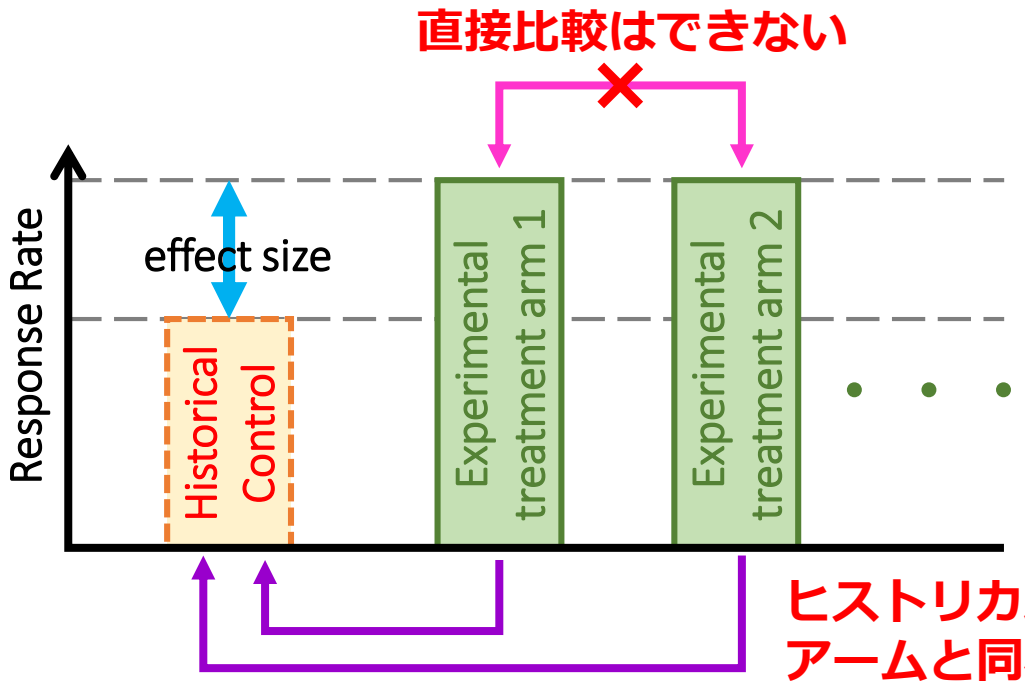
Green et al. (2002)では，上記デザインに加えて

- **ランダム化中止デザイン** (Randomized discontinuation design)

についても無作為化比較第II相試験のバリエーションとして加えている

1) Mandrekar, S.J. and Sargent, D.J. : Journal of Thoracic Oncology, 5(7), 932-934.

並行非比較レジメン (Parallel Noncomparative Regimens)

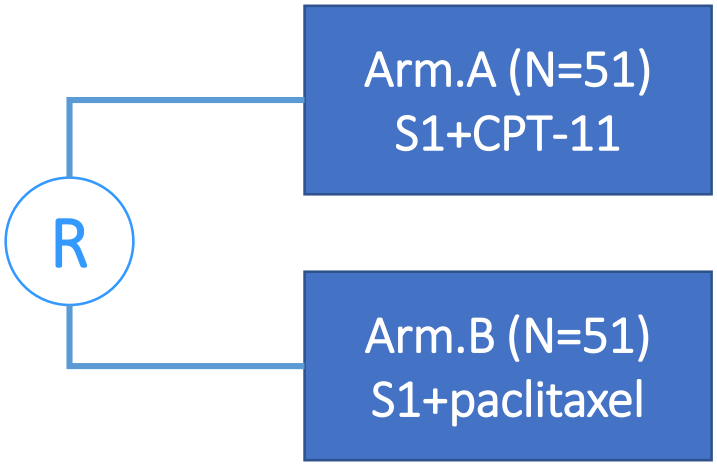


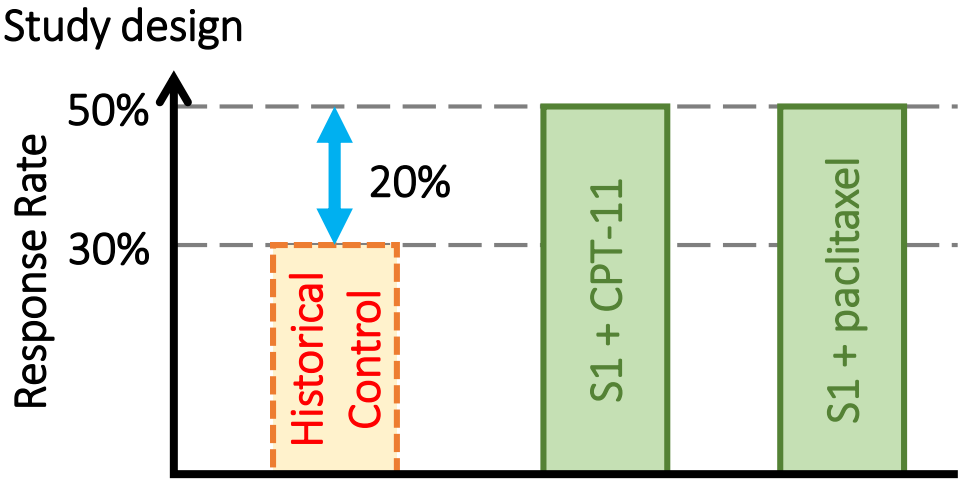
複数の試験治療アームのそれぞれに対して，ヒストリカル・コントロールとの比較を検討するデザイン.

- 複数の試験治療レジメンをランダム化することで，エンドポイントに対する，(1) 被験者選択バイアス，被験者背景のインバランスによる影響を排除したうえで評価できる.
- **試験治療アーム間の比較はできない(もし実施しても参考資料にしかならない).**
- 一方で，複数の試験治療レジメンがあり，それぞれに対しての直接比較を実施する前に有効性・安全性の情報が必要な場合(つまり，比較試験の前段階として)実施することは可能かもしれない.

並行非比較デザイン：事例

進行胃がんに対するS1+CPT-11 vs. S1+paclitaxel**の無作為化比較第II相試験**(OGSG0402)





母比率の検定においてα=0.05(two sided), 1-β=0.8となるように設定(n=50)

	No. of pts	Response					Response rate (%) (95%CI)
		CR	PR	SD	PD	NE	
S-1+ CPT-11	51	2	15	17	8	9	33.3% (20.8-47.9)
S-1+ paclitaxel	51	1	15	18	11	6	31.4% (19.1-45.9)

いずれの治療レジメンもヒストリカルコントロールの30%を有意に上回らなかった.

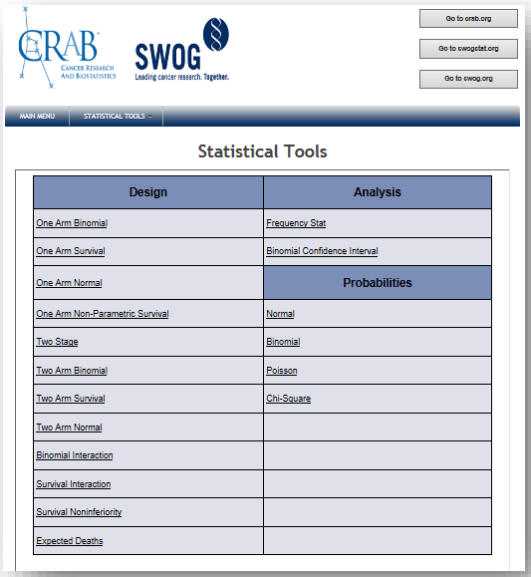
1) Sugimoto, N. et al.: Anticancer Research, 34, 851-858, 2014.

ランダム化非比較レジメンによる症例設計のサイト等紹介

ヒストリカルコントロールとの治療効果の差(effect size)がすべての試験治療で等しいとき，標本サイズは(**一つの比較での標本サイズ**)× (**試験治療群の数**) である.

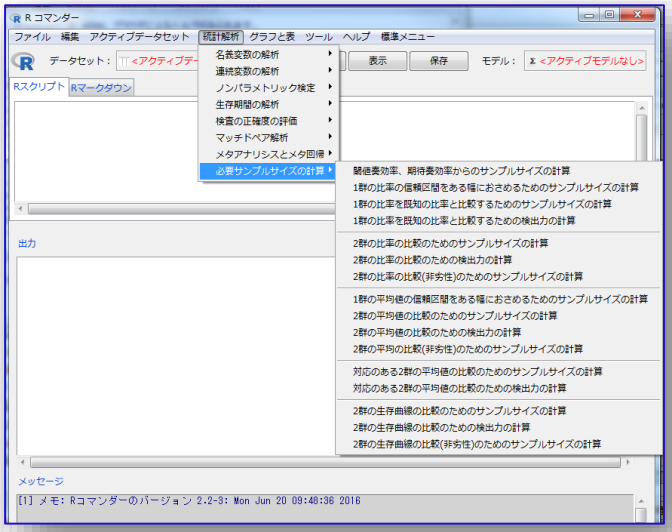
➡ **単アーム試験でのデザイン**(RR：One arm binomial，Survival：One arm survival)**に基づいて計算し，試験治療アーム数でかけ合わせればよい**

SWOGウェブサイト



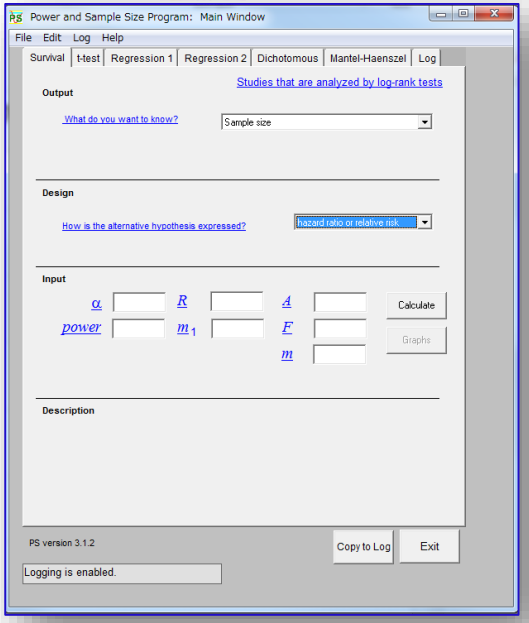
https://stattools.crab.org/

EZR



http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmed.html

PS



http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize

ランダム化選択デザイン(Randomized Selection Designs)

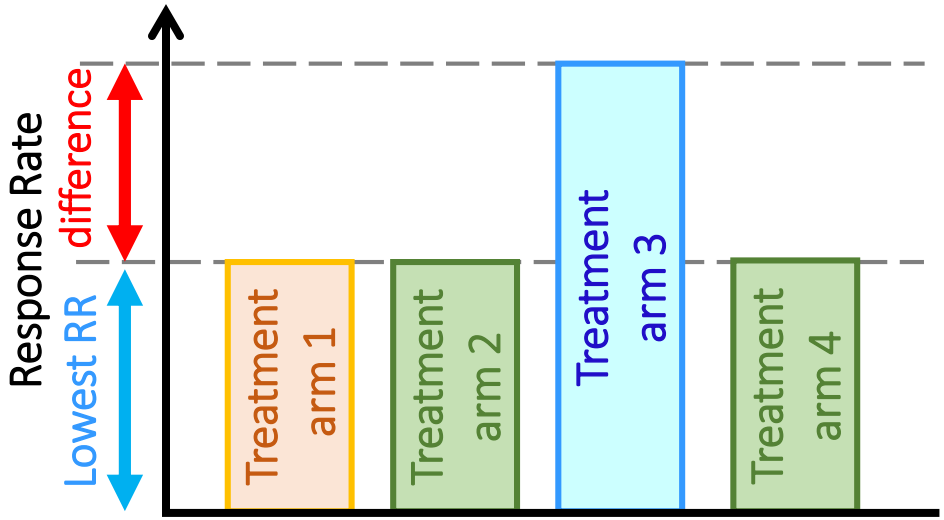
ランダム化非比較レジメン

複数の試験治療レジメンの中から，ヒストリカルコントロールとの比較を通じて候補となる試験治療を探す．

ランダム化選択デザイン(Simon et al., 1985)

複数の試験治療レジメンの中から，次の相に進めるものを”pick –the-winner”選択によって決定する．

Simon et al.: Cancer Treatment Report, 69, 1375-1381, 1985.



ランダム化選択デザインでは，第1種の過誤α，第2種の過誤βから症例数を決定するのではなく，

- 最小奏効率：R
- 最良治療レジメンとのRの差：D
- 最良治療レジメンの選択確率：P
- 試験治療レジメンの個数：K

に基づいて標本サイズが設定される．

ランダム化選択デザインとは，K-1個の治療レジメンが最小奏効率(最良以外は等しいと仮定)Rであり，最良治療レジメンはDの上乗せ効果があると仮定するとき，その最良治療レジメンを最良であると選択する確率をPで保証する標本サイズとして設定される．

ランダム化選択デザインの長所

選択確率P=90%のときの症例数(Simon et al., 1985)

RR (最良以外)	最良RR (D=15%)	1群当たりの症例数(Kは群数)		
		K=2	K=3	K=4
10%	25%	21	31	37
20%	35%	29	44	52
30%	45%	35	52	62
40%	55%	37	55	67
50%	65%	36	54	65
60%	75%	32	49	59
70%	85%	26	39	47
80%	95%	16	24	29

ランダム化選択デザインでは，試験治療レジメンの個数が増加しても，1群あたりの症例数は大きくは増加しない第III相試験の場合には，多重比較調整が必要なため，大幅に症例数が増加する).

試験治療アームのなかから第III相試験に進むべき唯一のもの(他の治療レジメンに比べて圧倒的に優れている)を選択するという意味では，ランダム化選択デザインは適している.

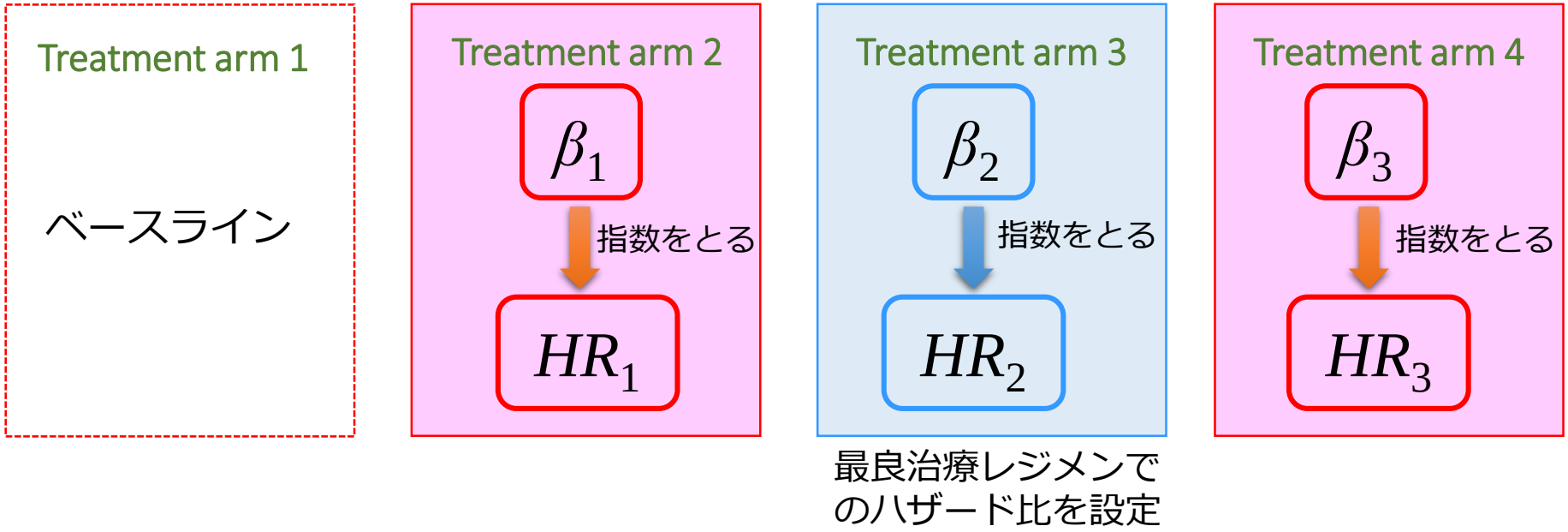
生存期間がアウトカムの場合のランダム化選択デザイン

Liu et al.(1993)は，ランダム化選択デザインを生存時間データに拡張している．

Liu et al.: Biometrics, 49, 391-398, 1993.

比例ハザードモデルで考える(治療数がK個では治療効果を表すパラメータはK-1個)

生存期間が最もPoorな治療レジメンをベースラインとして考える．



最良治療レジメンで
のハザード比を設定

生存時間がアウトカムの場合のランダム化選択デザインとは，K-2個の治療レジメンでのハザード比HR=1.0 (β_k=1)，最良治療レジメンのハザード比(or logrank検定の値)がHR(ex: HR=1.3)であると仮定するとき，その最良治療レジメンをPの確率で最良であると選択することを保証する標本サイズとして設定される．

予想したシナリオからズレた場合の最良レジメンの選択確率

シミュレーション回数3,000回

シナリオ1：RR=5%(最良レジメン以外)，最良RR=20%，　試験結果での差：15% (0%)

真のRR	5%/20%	10%/25%	15%/30%	20%/35%	30%/45%
K=2	0.91	0.96	0.95	0.94	0.91
K=3	0.89	0.95	0.92	0.90	0.87
K=4	0.90	0.92	0.90	0.88	0.84

シナリオ2：RR=10%(最良レジメン以外)，最良RR=30%，　試験結果での差：15% (-5%)

真のRR	5%/20%	10%/25%	15%/30%	20%/35%	30%/45%
K=2	0.71	0.86	0.90	0.89	0.88
K=3	0.69	0.85	0.87	0.83	0.82
K=4	0.71	0.85	0.85	0.80	0.80

シナリオ1は，最良レジメンとその他のあいだに15%のRRがあることを想定してデザインしたもとで，実際の試験において15%の差が認められた場合.
シナリオ2は，最良レジメンとその他のあいだに20%のRRがあることを想定してデザインしたもとで，実際の試験において15%の差が認められた場合.

- 想定された最良レジメンとその他のレジメンでのRRの差未満の結果しか得られない場合には，誤った治療レジメンを選択する確率が上昇する.
- 誤った治療レジメンを選択する確率は，真のRRが50%から遠くなるほど大きくなる傾向にある.

Crowley, J. and Hotering, A. (eds.): Handbook of Statistics in Clinical Oncology (3rd Edition), CRC Press, 2012.

ランダム化選択デザインでの欠点

ランダム化選択デザインでは，どの治療レジメンが優れているかといった情報は含まれていない．

本デザインでは，「**ある治療が非常に優れており，RRが他よりも15%上昇するときに，本試験で90%の確率で最も高いRRを示す**」ことを保証している．

したがって，ランダム化選択デザインでは，以下は保証されない．

n

■ **複数の有望な治療レジメンを選択する．**
　ランダム化選択デザインでは，有望な治療レジメンが複数である場合においても，(本試験で偶然に発生した)RRに基づいて一つを選択しなければならない．

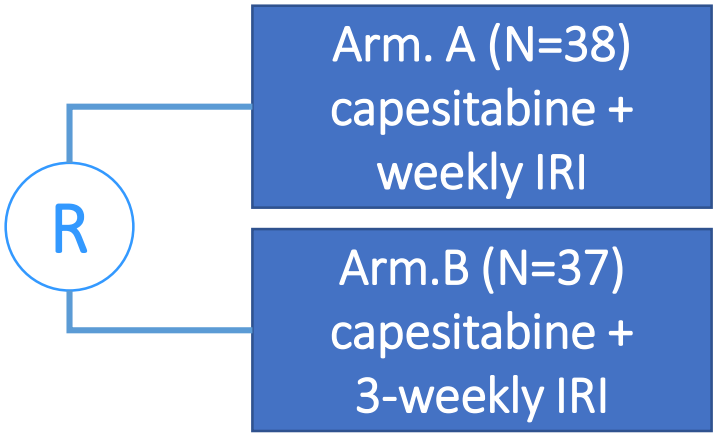
■ **どの治療法もそれほど違いがない場合でも一つを選択してしまう．**
　どの治療レジメンも治療効果の差Dを超えなくても治療レジメンを一つ決定しなければならない(Dの差に対する保証をしているわけではない)．つまり，どの治療レジメンも選ばないという選択肢はない．

統計学的評価は，各治療レジメンでのRR(or survival)の点推定値と95%C.I.によるものになるが，その結果をもって，上記問題に対処するしかない．

ランダム化選択デザインは，次の試験(例えば第III相試験)において，決定された有望な治療レジメンを検証する目的でない限り適用すべきでない(Green et al., 2012)

ランダム化選択デザイン：事例 (RRの場合)

転移性大腸がんに対する1st lineとしてのcapesitabineと2種類のirinotecan投与に関するランダム化比較第II相試験



本試験では，いずれかの治療レジメンが15%以上のRRの差が出たときに，90%の選択確率で有望な治療レジメンを選択するように設定している．

論文中では，”no formal statistical comparisons”と記載されており，群間の比較(ex: Fisherの正確検定)は行わず，それぞれの奏効率に対して，Clopper-Pearsonの正確な信頼区間を計算することのみがStatistical Analysisに記載されている．

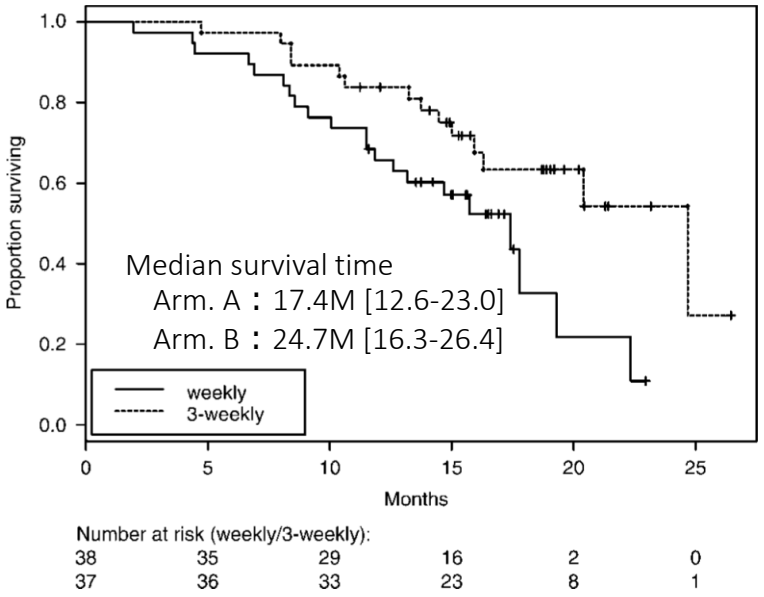
奏効率RR

Arm. A： 18% [95% C.I. 8-34%]

Arm. B： 35% [95% C.I. 20-53%]

Arm. Bのほうが有効

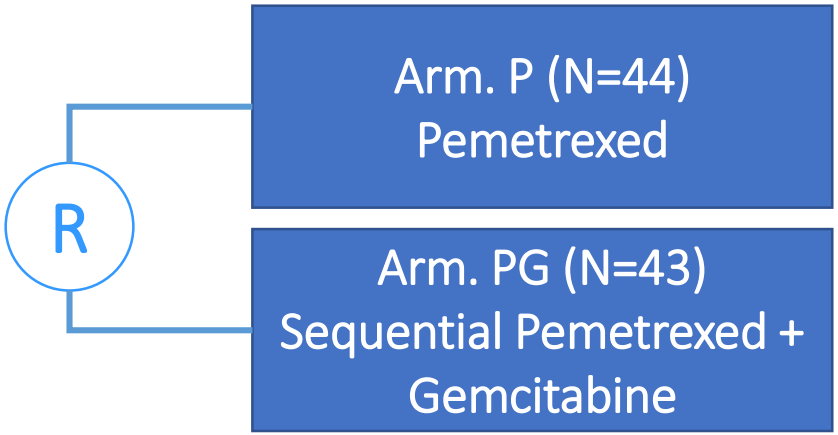
Arm.Bのほうが奏効率，生存期間ともに長いことから，CAP+IRI併用療法において，IRIの投与レジメンは3-weeklyのほうが良いかもしれない．



Borner, M.M. et al.: Annals of Oncology,16, 282-288, 2005.

ランダム化選択デザイン：事例 (Survivalの場合)

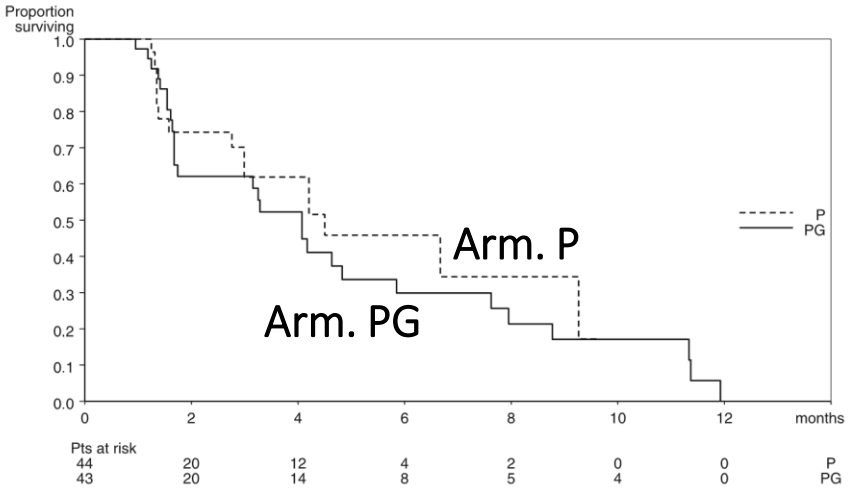
高齢者あるいはプラチナ不適進行非小細胞肺がん患者に対する1st lineとしての
Pemetrexed vs. Sequential Pemetrexed+Gemcitabine**療法のランダム化比較第II相試験**



本試験では，無増悪生存期間(PFS)において，ハザード比が0.82(PFSが長いArm / PFSが短いArm)となるとき，80%の確率でPFSが長いArmを選択するように設定されている.

PFSが長いArmでの想定MST：9M
PFSが短いArmでの想定MST：5M

どちらのArmのMSTであるかは考慮されない(考慮できない)



Arm PでのMST = 3.3M (95%CI: 2.0 – 4.4)
ARM PGでのMST = 3.3M (95%CI: 1.7-4.1)

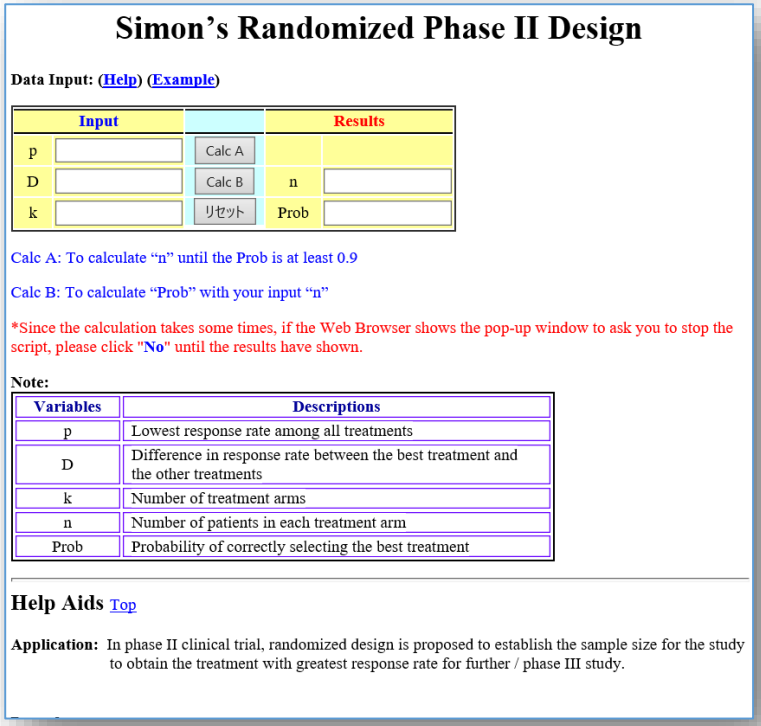
両群の違いは認められなかった (本論文においても，ログランク検定等の比較は行われていない).

選択デザインでは，群間差を比較するような統計手法がなく，「どのような状況になるとpositiveなのか」を決定することができない(論文によっては検定を実施しているが，あくまでも参考資料程度の結果である).

ランダム化選択デザインによる症例設計のサイト紹介

香港中文大学

Centre for Clinical Research and Biostatistics HP



Simon's Randomized Phase II Design

Data Input: (Help) (Example)

Input			Results	
p		Calc A		
D		Calc B	n	
k		リセット	Prob	

Calc A: To calculate "n" until the Prob is at least 0.9

Calc B: To calculate "Prob" with your input "n"

*Since the calculation takes some times, if the Web Browser shows the pop-up window to ask you to stop the script, please click "No" until the results have shown.

Note:

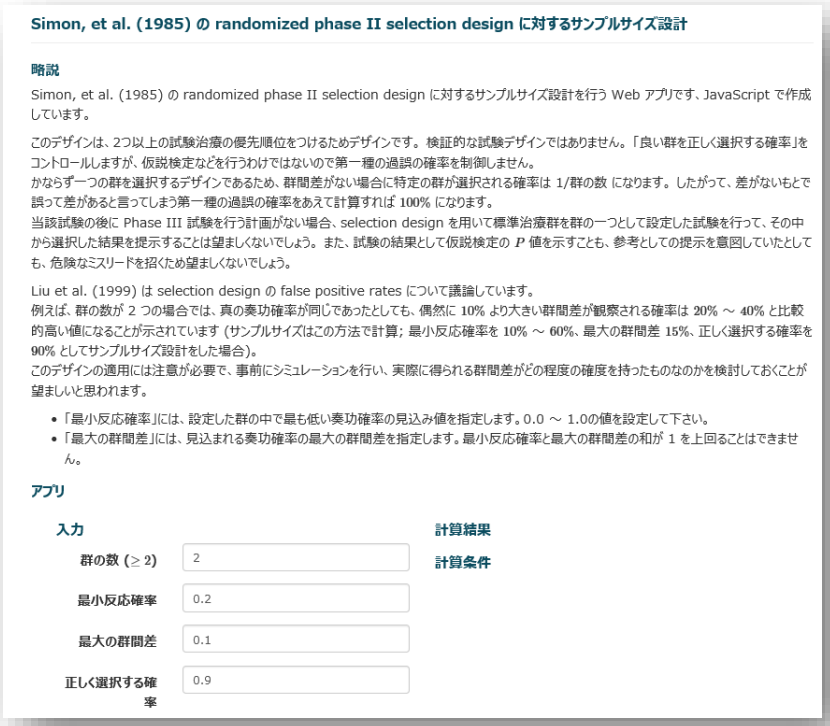
Variables	Descriptions
p	Lowest response rate among all treatments
D	Difference in response rate between the best treatment and the other treatments
k	Number of treatment arms
n	Number of patients in each treatment arm
Prob	Probability of correctly selecting the best treatment

Help Aids Top

Application: In phase II clinical trial, randomized design is proposed to establish the sample size for the study to obtain the treatment with greatest response rate for further / phase III study.

https://www2.ccrb.cuhk.edu.hk/stat/phase2/Randomized.htm

千葉大学 グローバル臨床試験学 HP



Simon, et al. (1985) の randomized phase II selection design に対するサンプルサイズ設計

暗説

Simon, et al. (1985) の randomized phase II selection design に対するサンプルサイズ設計を行う Web アプリです、JavaScript で作成しています。

このデザインは、2つ以上の試験治療の優先順位をつけるためデザインです。検証的な試験デザインではありません。「良い群を正しく選択する確率」をコントロールしますが、仮説検定などを行うわけではないので第一種の過誤の確率を制御しません。
かならず一つの群を選択するデザインであるため、群間差がない場合に特定の群が選択される確率は 1/群の数 になります。したがって、差がないもとで誤って差があると言ってしまう第一種の過誤の確率をあえて計算すれば 100% になります。
当該試験の後に Phase III 試験を行う計画がない場合、selection design を用いて標準治療群を群の一つとして設定した試験を行って、その中から選択した結果を提示することは望ましくないでしょう。また、試験の結果として仮説検定の P 値を示すことも、参考としての提示を意図していたとしても、危険なミスリードを招くため望ましくないでしょう。

Liu et al. (1999) は selection design の false positive rates について議論しています。
例えば、群の数が 2 つの場合では、真の奏功確率が同じであったとしても、偶然に 10% より大きい群間差が観察される確率は 20% ～ 40% と比較的高い値になることが示されています (サンプルサイズはこの方法で計算; 最小反応確率を 10% ～ 60%、最大の群間差 15%、正しく選択する確率を 90% としてサンプルサイズ設計をした場合)。
このデザインの適用には注意が必要で、事前にシミュレーションを行い、実際に得られる群間差がどの程度の確度を持ったものなのかを検討しておくことが望ましいと思われます。

- 「最小反応確率」には、設定した群の中で最も低い奏功確率の見込み値を指定します。0.0 ～ 1.0の値を設定して下さい。
- 「最大の群間差」には、見込まれる奏功確率の最大の群間差を指定します。最小反応確率と最大の群間差の和が 1 を上回ることはできません。

アプリ

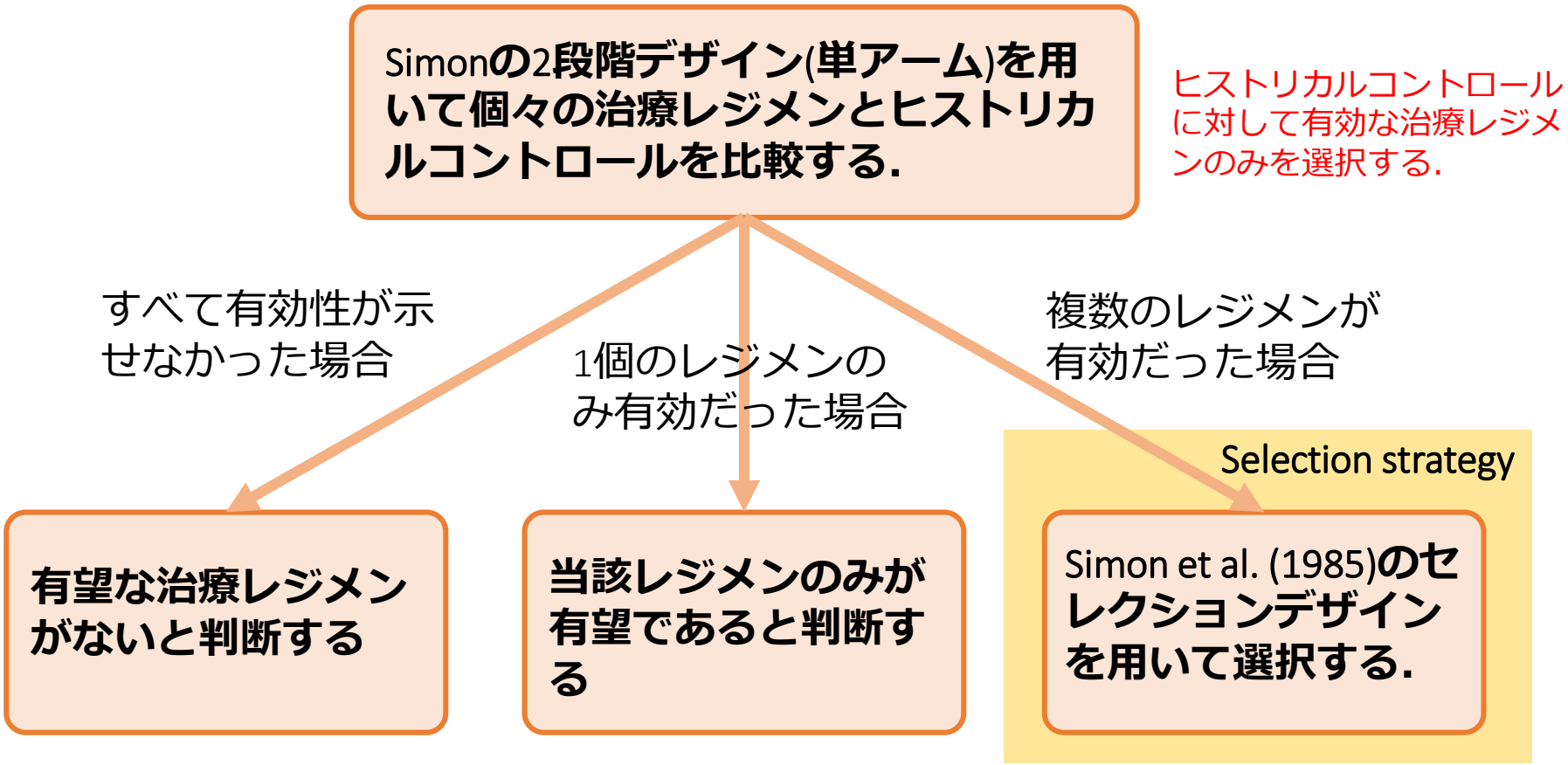
入力		計算結果
群の数 (≥ 2)	2	計算条件
最小反応確率	0.2	
最大の群間差	0.1	
正しく選択する確率	0.9	

http://nshi.jp/contents/js/selection/

あとは，Rのパッケージclinfunのなかの関数pselectを用いることで選択確率が計算できるため，それを応用すれば症例設計ができる.

Screened selection design(Yap, Pettitt and Billingham, 2013)

- Simon et al.(1985)のランダム化選択デザインでは，標準治療に対して，有効な治療か否かという情報は含まれていない(新規治療候補の比較).
- Simon et al.(1985)のランダム化選択デザインでは，どの治療レジメンも選ばないという選択肢はない.



症例数は，(2段階デザインでの症例数)×群数のセレクションデザインの症例数のいずれか多いほうを選択する.

Yap, C., Pettitt, A. and Billingham, L.: BMC Medical Research Methodology, 13:87, 2013.

Screened selection designの事例

局所進行直腸癌に対する術後化学放射線療法(CAPOX+放射線療法)と術後化学療法(CAPOX)に対するランダム化比較第II相試験

主表評価項目：病理学的完全奏効率：pCR

各群に対してSimonの2段階デザインを適用

CAPOX+Rad**群**：Capecitabine+Oxaliplatin+Radiation

CAPOX**群**：Capecitabine+Oxaliplatin

Simonの2段階デザイン

閾値pCR=10%, 期待pCR=25%，α=0.05, β=0.10

症例数：49 (per arm)

24例で中間解析

→ 1例以下の場合には早期無効中止

すべて有効性が示せなかった場合

1個のレジメンのみ有効だった場合

複数のレジメンが有効だった場合

有望な治療レジメンがないと判断する

当該レジメンのみが有望であると判断する

Simon et al. (1985)のセレクションデザインを用いて選択する.

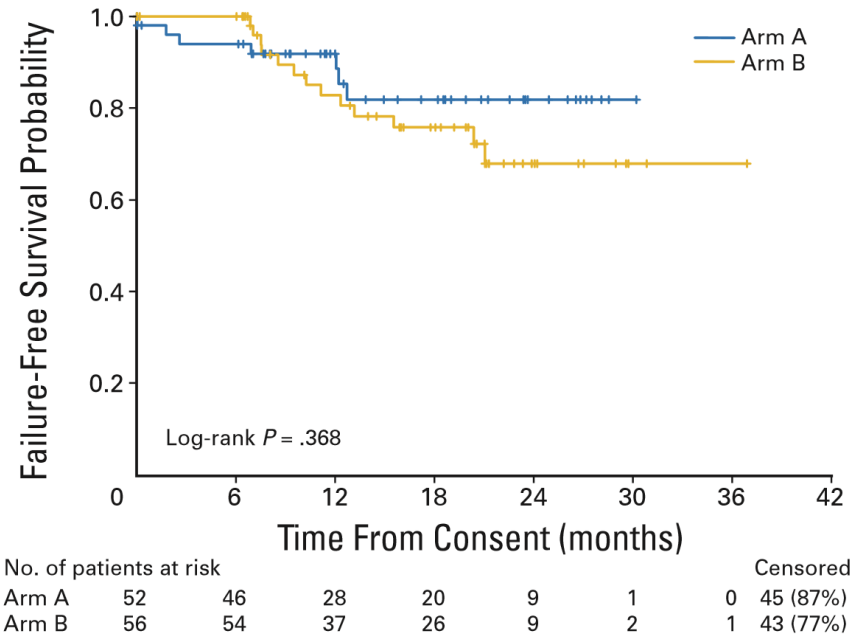
治療効果の差：15%，選択確率：90%

両方の治療レジメンの間に大きな差異が認められない場合は，他の要因(feasibility, toxicity)を考慮して推奨レジメンを選択する.

Fernandez-Martos, C. et al.: Journal of Clinical Oncology, 28, 859-865, 2010.

続き

End Point	Arm A: Post-operative Adjuvant CT (n = 52) No.	%	Arm B: Induction CT (n = 56) No.	%	P*
pCR	7	13	8	14	.94
95% CI, %	5.6 to 25.8		6.4 to 26.2		
Downstaging	30	58	24	43	.13
95% CI, %	43.2 to 71.3		29.7 to 56.8		
R0 resection rates	45	87	48	86	.40
TRG†					
4: complete regression	7	15	8	15	.88
3: > 50% of tumor mass	22	48	20	37	
2: ≥ 25%-50% of tumor mass	11	24	13	24	
1: < 25% of tumor mass	2	4	3	6	
0: no regression	1	2	3	6	
Not otherwise specified	3	7	7	13	



いずれの群も閾値pCR(10%)対して有意な差は認めらない.

Arm.A(p=0.261(one sided)), Arm.B(p=0.193)

シミュレーションによる評価結果

ヒストリカルコントロール = 0.05, シミュレーション回数10,000回

真の有効率 (PA-PB)	SSD			Modified SSD			SWE		
	Arm A	Arm B	No Arm	Arm A	Arm B	No Arm	Arm A	Arm B	No Arm
(0.01,0.01)	0.025	0.025	0.950	0.025	0.025	0.950	0.500	0.500	0.000
(0.1,0.1)	0.455	0.454	0.091	0.311	0.311	0.378	0.500	0.500	0.000
(0.2,0.2)	0.500	0.498	0.002	0.320	0.320	0.360	0.500	0.500	0.000
(0.3,0.3)	0.500	0.500	0.000	0.334	0.335	0.331	0.500	0.500	0.000
(0.01,0.03)	0.023	0.167	0.810	0.021	0.164	0.815	0.315	0.685	0.000
(0.01,0.2)	0.002	0.950	0.048	0.001	0.947	0.052	0.003	0.997	0.000
(0.20,0.35)	0.100	0.900	0.000	0.042	0.805	0.153	0.099	0.901	0.000
(0.2,0.4)	0.047	0.953	0.000	0.017	0.897	0.086	0.046	0.954	0.000

＊SSD：Screened selection design, SWD：Simonのランダム化選択デザイン

Modified SSDは，Selection designのなかに閾値を設定し，閾値以上のなかで最大の有効性(RR or HR)を示すArmを選択する方法である．

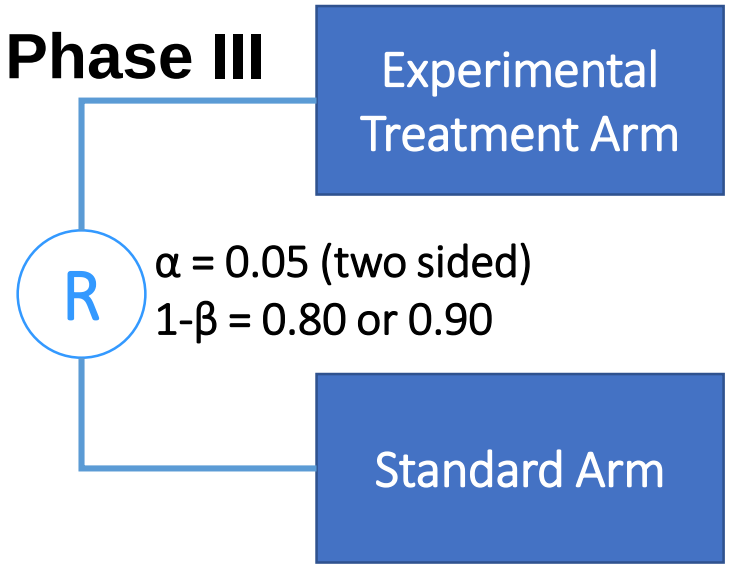
Simonの2段階デザインを前評価に利用することで，既存治療と同等(あるいは以下)の候補レジメンのスクリーニングを行ったうえで最良レジメンを選択できることがわかる．

ランダム化スクリーニングデザインの説明の前に：復習

	検定により評価	標本サイズで規定
	H_0が正しい [H_1：偽]	H_1が正しい [H_0：偽]
H_0を棄却 [H_1を受容]	第1種の過誤α (H_0が正しいのに棄却)	検出力1-β (H_0が正しいので受容)
H_0を受容 [H_1を受容]	正しい判断1-α (H_1が正しいので棄却)	第2種の過誤β (H_1が正しいのに受容)

- **第2種の過誤が一定水準未満(β未満)になるように，標本サイズを規定する** (本当はpositiveなのに本試験を通じてnegativeと判断する可能性を低くする).
- **統計解析の結果，第1種の過誤が一定水準未満(α未満)であるか否かを確認する**(本当はnegativeなのに本試験結果を通じてpositiveと結論付けるエラーがかなり低いことを確認する).

ランダム化スクリーニングデザイン(1)　(Randomized Screening Designs)



Study designのために必要な情報

- 治療効果の差(Effect size)：HR
  → **評価の指標なので変更不可**
- 第1種の過誤[有意水準](αエラー)
  → **単アーム試験と同等(0.1)，あるいは，大きく設定**
- 第2種の過誤(βエラー)
  → **標本サイズでのみ規定できるため，変更は難しい.**

Rubinstein et al.(2005)は，有意水準をSWOGの単アーム試験と同様の水準(有意水準15～20%(one sided)，検出力90%)とすることで，標本サイズを比較的小さくすることができることを指摘している．これをランダム化スクリーニングデザインという.

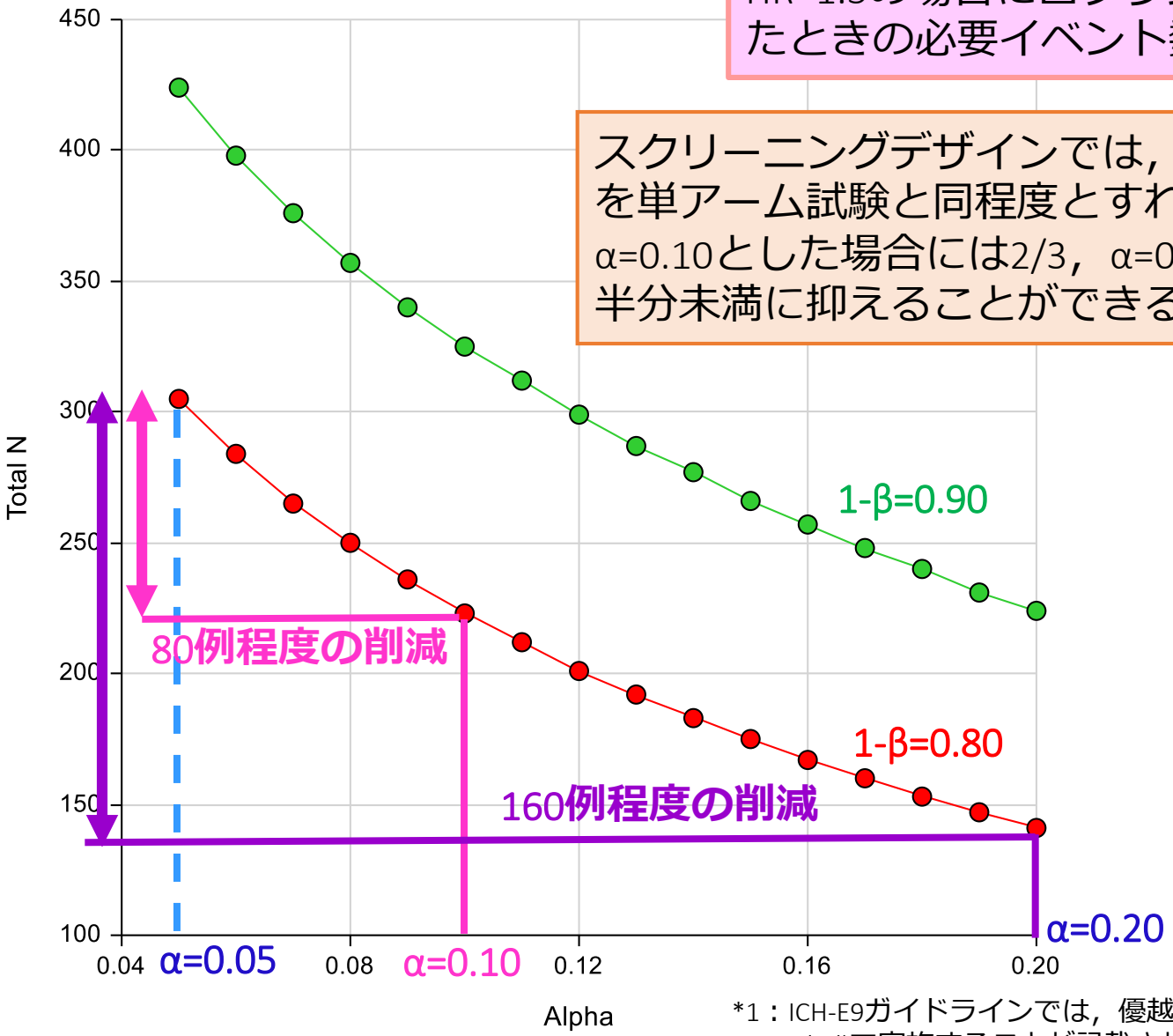
Rubinstein, L.V. et al: Journal of Clinical Oncology, 23, 7199-7206, 2005.

αエラーを増加させると必要症例数がどの程度変化するのか？

PIIIと同水準*1

HR=1.5の場合にログランク検定で評価したときの必要イベント数(total)

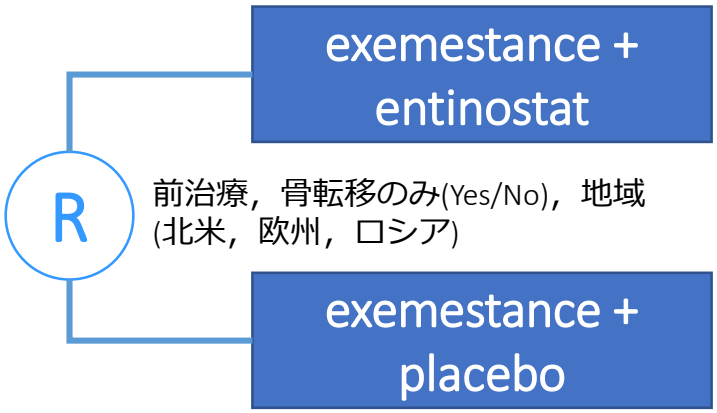
スクリーニングデザインでは，αエラー，βエラーを単アーム試験と同程度とすれば，必要症例数をα=0.10とした場合には2/3，α=0.20とした場合には半分未満に抑えることができる．



*1：ICH-E9ガイドラインでは，優越性試験の場合にはα=0.025(one sided)で実施することが記載されている．

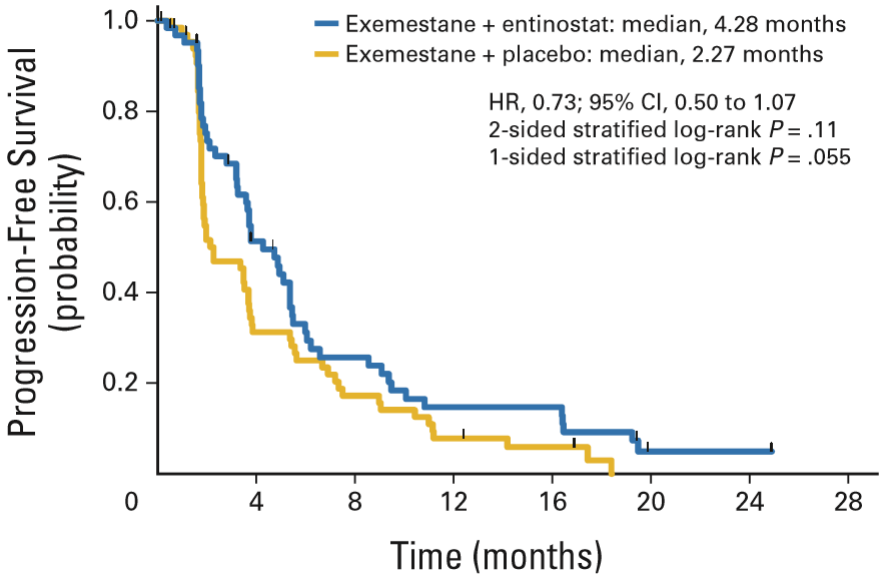
ランダム化スクリーニングデザイン：事例

局所再発／転移性の閉経後エストロゲン受容体(Estrogen Receptor)陽性進行乳がん患者に対するEntinostatに関する無作為化比較第II相試験



既存研究において，閉経後進行乳がん患者に対するアロマターゼ阻害薬でのPFSのMSTは3.7ヵ月であることが報告されている．この研究では，アロマターゼ阻害薬に対する治療抵抗化の予防・低減が期待できるeitinostatを投与することで，PFSが2.3ヵ月(PFS=6ヵ月)伸びることを期待している(HR=0.62)．

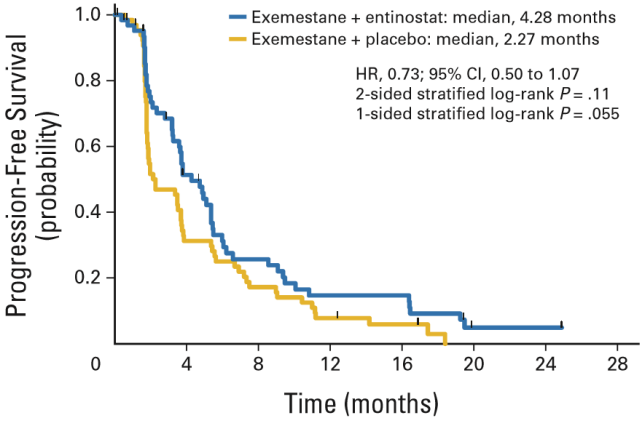
上記設定のもとで，ログランク検定(one sided α=0.10)，を実施するとき，検出力1-βが90%となる必要イベント数は112例(total)である．



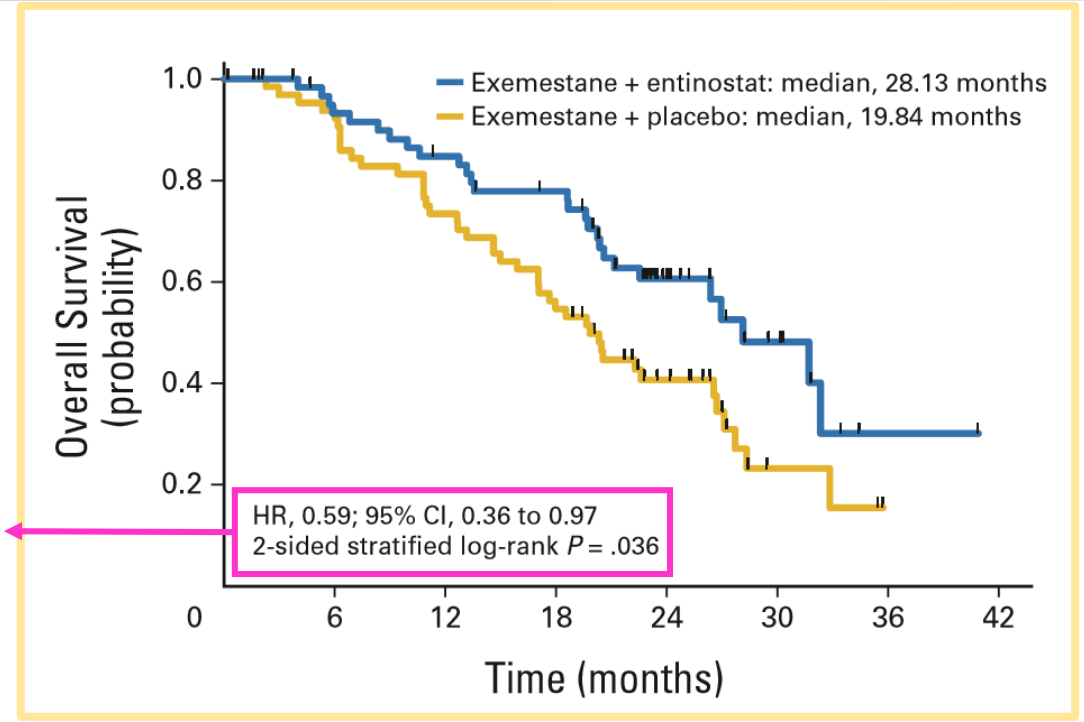
PFSの片側p値が0.055だった．本試験ではα=0.10なので，positiveな結果が得られた．すなわち，eitinostat投与はアロマターゼ阻害薬の効果を伸ばすことが期待できる．

Yardley, D.. et al: Journal of Clinical Oncology, 31, 2128-2136, 2013.

先ほどの事例の続き



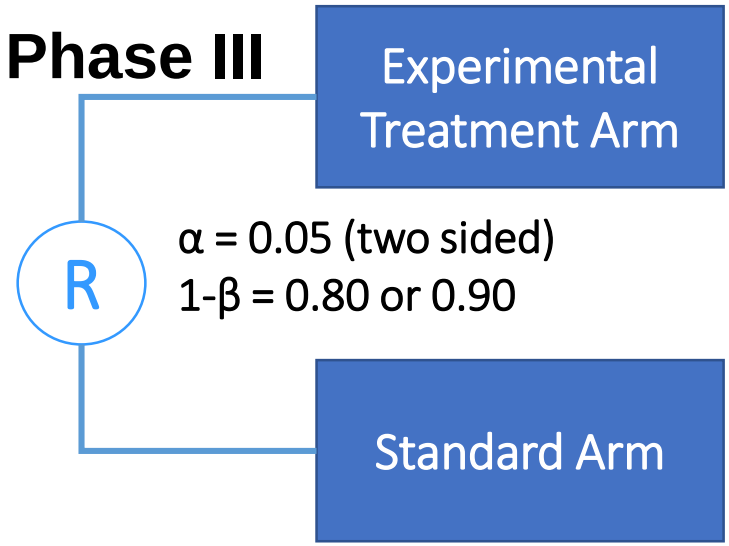
OSが層別化ログランク(コクラン・マンテルヘンツェル検定)において p=0.036(two sided)だった.



第III相試験は必要ないのではないか？

本試験は，eitinostat投与によってアロマターゼ阻害薬に対する治療抵抗化の予防・低減ができるかを検討する試験であり，OSは副次的評価項目に過ぎない．したがって，**本治療レジメンによる予後への影響については，第III相試験で確定させる必要がある**(ただし，OSが大幅に伸びることを示せたのは，RPII試験を実施したからである．).

ランダム化スクリーニングデザイン(2)　(Randomized Screening Designs)



Study designのために必要な情報

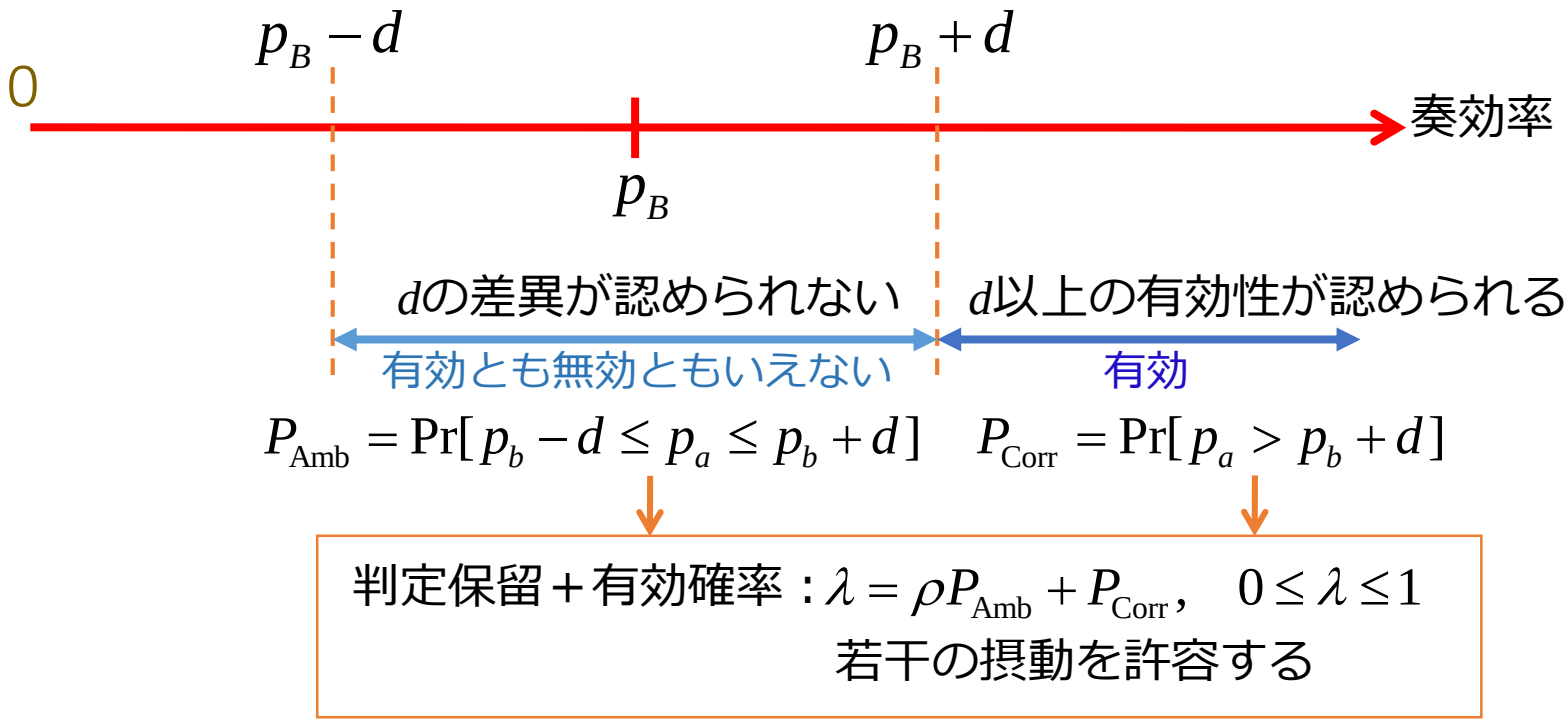
- 治療効果の差(Effect size)：HR
  → **判定保留を考慮して甘めにする**
- 第1種の過誤[有意水準](αエラー)
  → **変更しない**
- 第2種の過誤(βエラー)
  → **標本サイズでのみ規定できるため，変更は難しい.**

Sargent & Goldberg(2005)は，通常のエフェクトサイズに対して，判定保留の領域を考慮することで，その基準を甘めに設定することにより，症例数を少なくする方法を提案している.

Sargent, D.J. & Goldberg, R.M.：Statist. Med. 20, 1051-1060, 2001

判定の方法

奏効率の仮定：A群＞B群(多群の場合はB群の位置が増える)



対照が存在しない場合(pairwise**比較**)

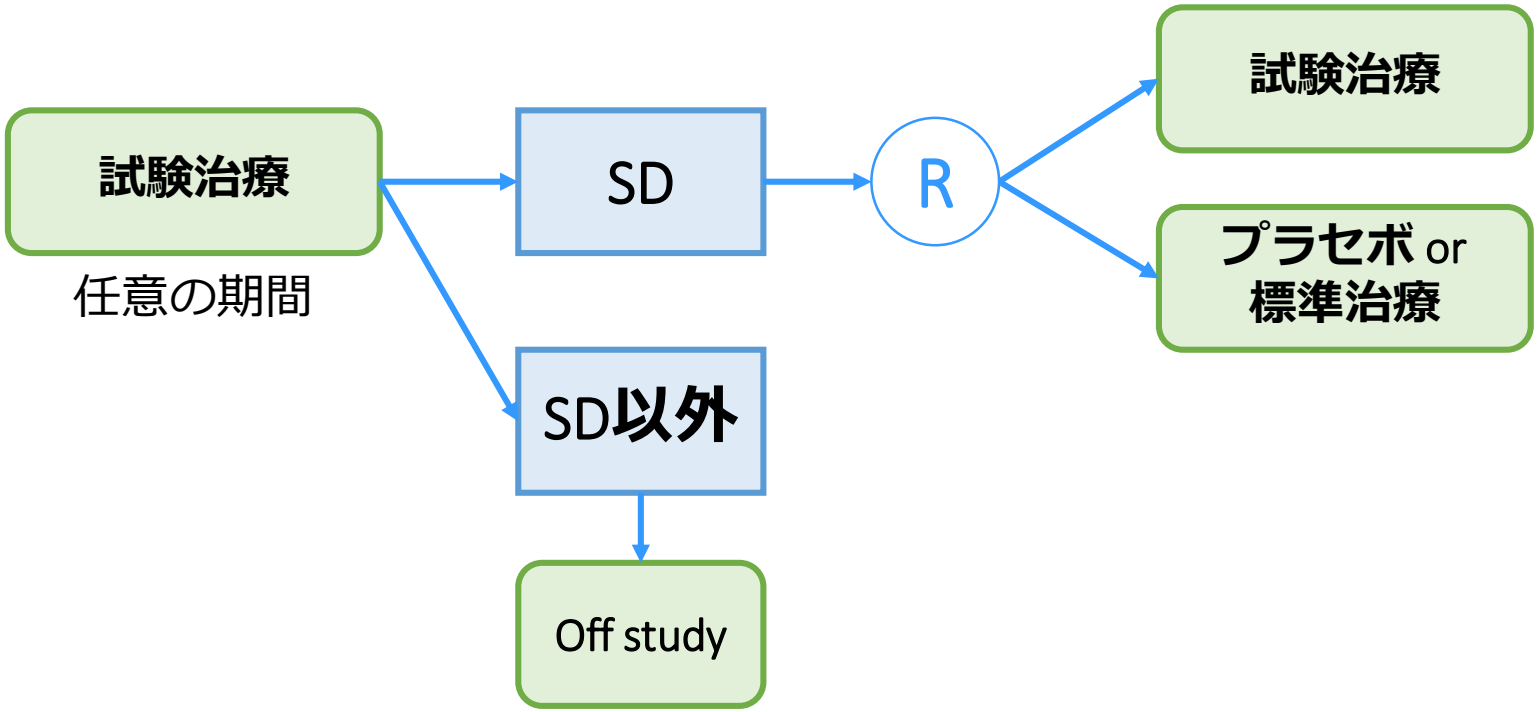
摂動を許容したもとで，1個以上の治療法が他群よりも有意になるように設定

対照が存在する場合(閉検定手順)

摂動を許容したもとで，1個以上の治療法が対照群よりも有意になるように設定

Sargent, D.J. & Goldberg, R.M.：Statist. Med. 20, 1051-1060, 2001

ランダム化中止デザイン(Randomized Discontinuation Designs)



○ 著効例，無効例(急速に進行する集団)が除外されるため，ランダム化される集団が均一になる．

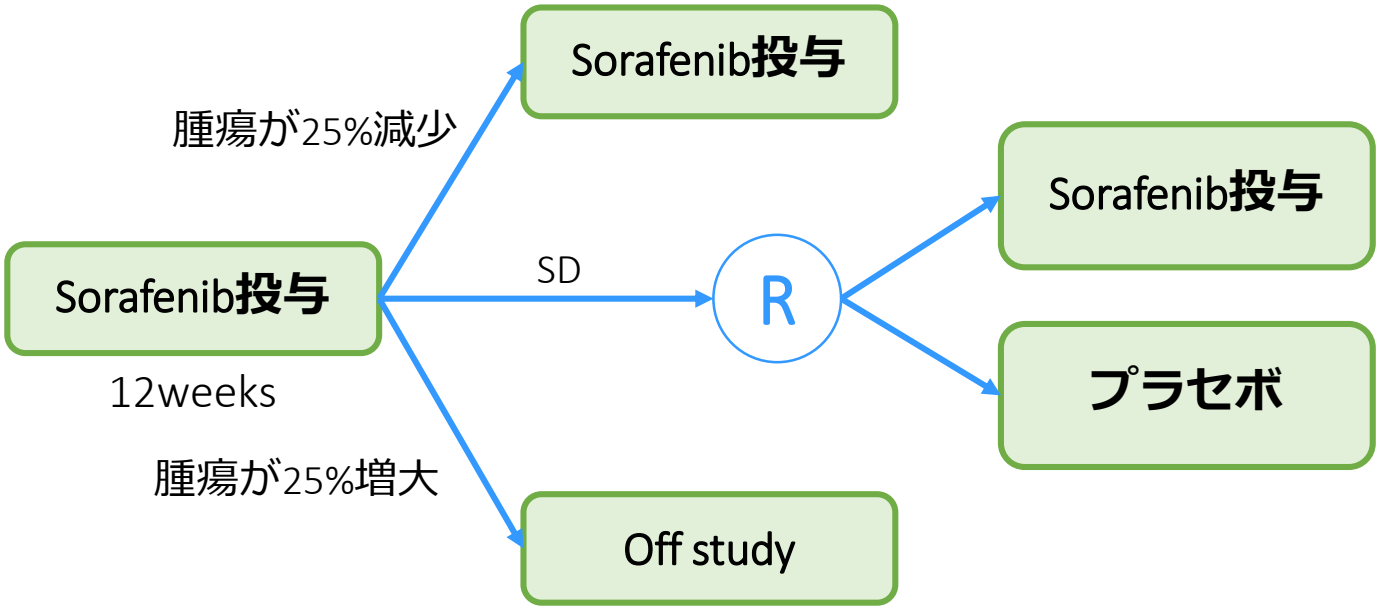
× プラセボ(or 標準治療)群に対して，試験治療を実施しており，正当な評価にならない可能性がある．

× SD以外の症例を見込まなければならないため，全体としての症例数は，通常の試験デザインに比べて多くなる可能性がある(Fredilin & Simon, 2005)．

Fredilin, B. and Simon, R.: Journal of Clinical Oncology, 5094-5095, 2005

ランダム化中止デザイン：事例

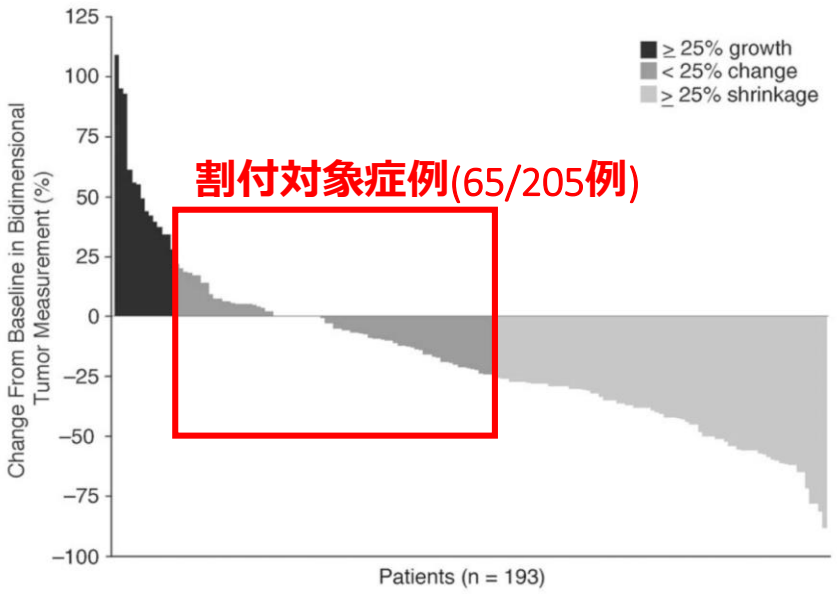
転移性腎細胞がん患者に対するSorafenib療法のプラセボ対照ランダム化中止試験

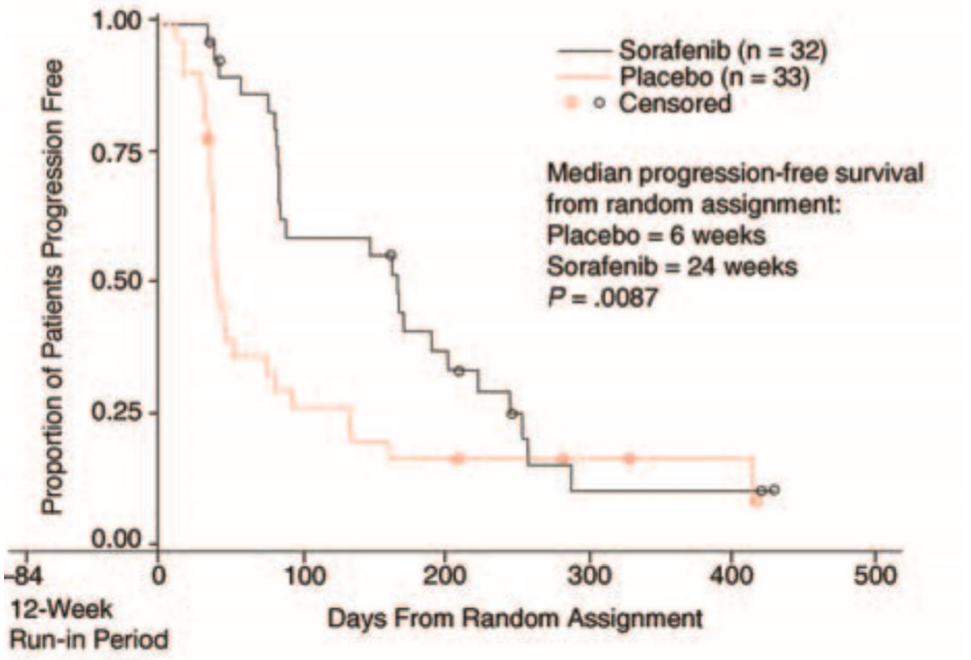


ランダム化中止デザインでは，全例が投与されたSorafenib投与でのSD症例の割合を計算しなければならない．この事例では，シミュレーション(PFSではなく腫瘍増加率)を通して，全登録例の43%がSD症例になることをシミュレーションにより計算している．

そして，割付12week後のSorafinib投与群とプラセボ投与群の無増悪生存率を90%,70%としたもとで，Sorafinibに対する選択確率を80%となるように症例設計を行なった．その結果，必要例数が65例，全登録例数が144例となった．

事例の結果



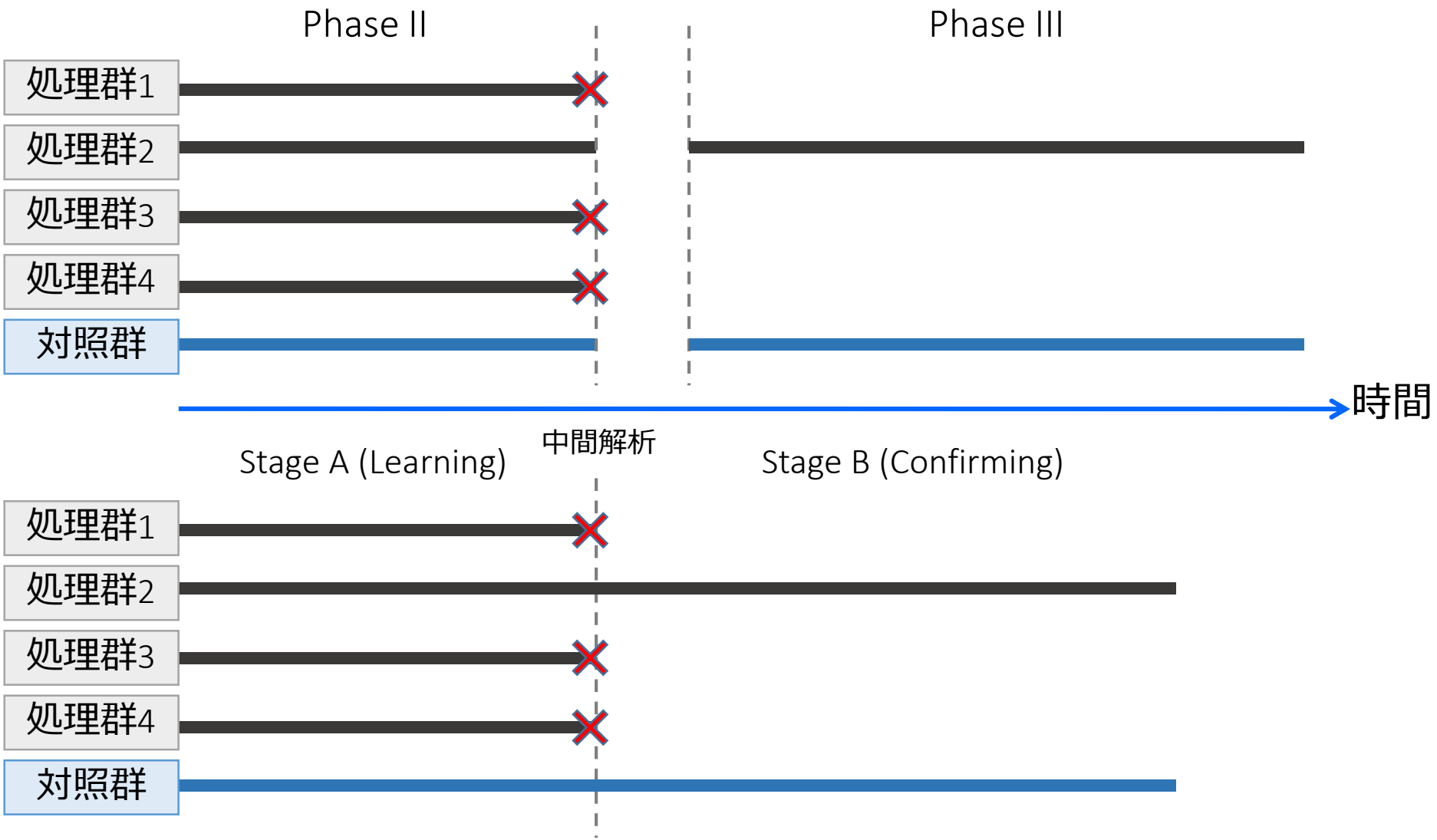


Sorafenib投与を投与することで，Placeboに比べてPFSが大幅に伸びた.

しかしながら，PR症例も多いことから，ランダム化中止デザインにする必要があったのかは不明(例えば，Screening designでも同様の結果が得られ，かつより少ない症例数で実施が可能だったはず).

最近の潮流：Seamless II/IIIデザイン(ひとつの例)

・ステイジ1でレジメンのスクリーニング(閉検定手順)，ステイジ2で対照群との比較
・StageAにおいて，OSについて著効な群がある場合には，早期有効中止



無作為化比較第II相試験でのエンドポイント：OS vs. PFS

Rubinstein et al.(2009)は，無作為化比較第II相試験では，OSよりもPFSをサン推奨している．

- Time-to-progressionのほうがTime-to-deathよりも短いことから，早期に生存期間をconfirmできる．
- 1st lineでは，OSでは後治療の影響を受けるため，当該治療の評価にならない可能性がある(PFSではその影響は少ない)．
- 上記理由から，PFSのほうがOSにくらべてHRが大きく，多くの症例数を必要とするかもしれない．

一方で，PFSの場合には，観測バイアス(検査日／診断日，測定間隔)，報告バイアス(画像評価に対する主観的バイアス)，患者脱落バイアスなどのバイアスに注意する必要がある．とくに，進行膵癌のような予後不良な癌の場合は深刻である．

PII studyという特性を考えればPFSを選択することが一つではあるものの，RECISTが利用できない状況では利用できない．また，予後不良がんの場合には，PFSでなくてもよいと考えられる．

Rubinstein L.V., et al. : Clinical Cancer Research, 15, 1883-1890, 2009.

第II相試験における単アーム試験と多アーム試験の選択

十分な被験者を確保できるか？(>100)	適切な標準治療が存在するか？	RRによる評価が可能か？	研究母集団対象のprognostic heterogeneity	Combination therapyか否か？	適切なヒストリカルデータが存在するか？	Single Arm	Randomized	コメント
―	±	±	±	±	±	＋＋＋	―	患者リクルートが限定された試験
＋	―	±	―	―	＋	＋＋	＋	標準治療が存在しない(患者母集団が比較的多い)重症患者に対する単剤投与試験[フィージビリティ等]
＋	＋	±	―	―	＋	＋＋	＋＋	標準治療が存在する(患者母集団が比較的多い)2nd line以降の患者に対する単剤投与試験
＋	±	―	―	±	＋	＋	＋＋	RECIST評価が困難であるかTime-to-Eventが利用される試験
＋	＋	±	―	＋	＋	＋	＋＋	転移性乳がん・大腸がんの1st line治療としての上乗せ効果に関する試験
＋	±	±	＋	±	＋	＋	＋＋＋	治療感度が臨床的・病理学的等によって変化する疾患部位に対する試験
＋	±	±	±	＋	―	―	＋＋＋	適切なヒストリカルデータが存在しない試験

Gan, H.K., Grothey, A. and Pond, G.R. et al.: Journal of Clinical Oncology, 28, 2641-2647, 2010.より一部修正

結びに代えて

単群試験

- **信頼できる標準治療に対するヒストリカルコントロールが存在する場合**
  Taylor et al. (2006)[1)]はRCTにおいて問題になるのは試験群のバイアスであることを指摘している(つまり，信頼がおけるヒストリカルコントロールがある場合には単群試験のほうが効率が良い).
- **新治療の治療完遂率を評価する場合**
- **少数例の患者を研究対象とする場合**

無作為化比較試験

- **標準治療が存在しない場合.**
- **標準治療が急激に変化したことで信頼できるヒストリカルコントロールが存在しない場合**
- **対象患者が変化した場合(分子標的薬で行われるサブセット等)**

抗がん剤治療に対するstrategyが急激に変化している現在において，単群試験 vs 無作為化比較試験ではなく，相互補完しながら適材適所で使うことが重要であり，第III相試験が難しいから無作為化比較第II相試験で実施するというのは誤り．結局は，第III相試験でないと試験レジメンの有効性・安全性はconfirmできない.

ご清聴ありがとうございました