Biostatistics Seminar 2019

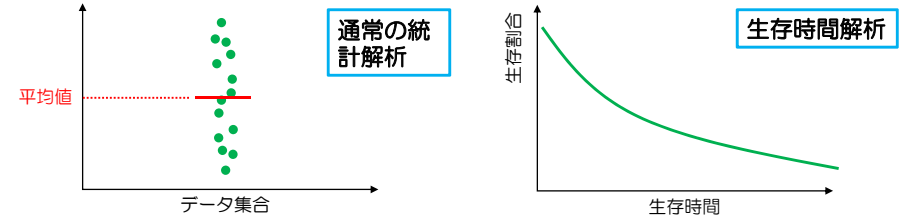
医学統計セミナー

- 生存時間解析(1) -
生存曲線の推定と比較

下川敏雄
和歌山県立医科大学附属病院 臨床研究センター

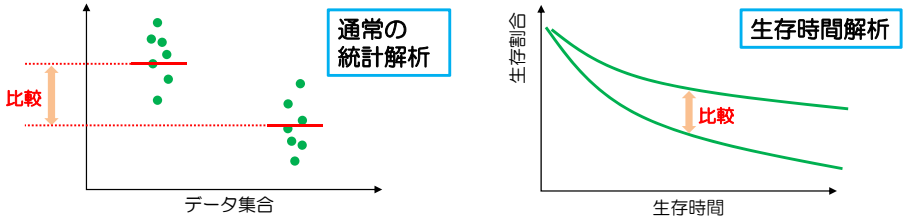
通常の統計解析と生存時間解析の違い

■ データ集合の要約について考えた場合



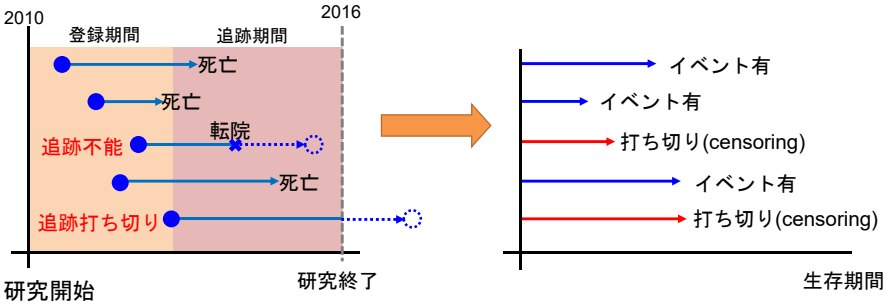
平均値などデータ集合の代表値を推定する　　生存曲線を推定する

■ 群間比較を考えた場合



代表値(例：平均値)を比較する　　生存曲線を比較する

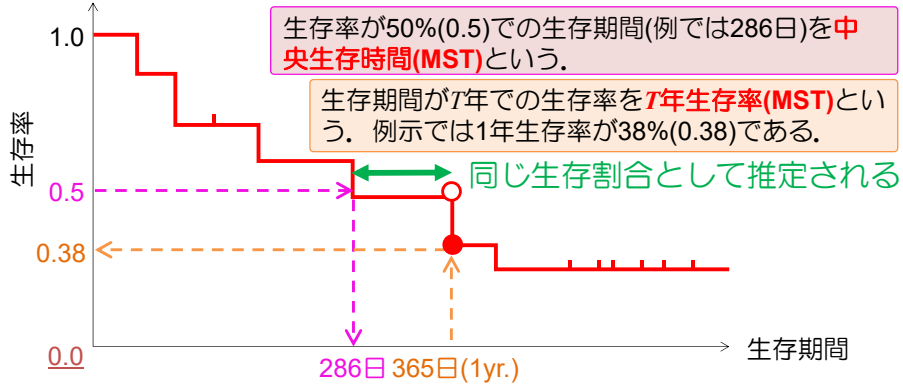
生存期間データのメカニズム：打ち切り(censoring)



生存時間データは，個体の死亡(あるいは故障)までの時間を観測して得られる．観測される生存時間は生存時間(確率変数)Tの実現値tであるか，あるいは中途打ち切りが生じたときには，その打ち切り(censoring)時点の値である．

本報告でとり扱う右側打ち切りとは，個体の登録時点はわかるものの，追跡期間中に死亡(event)が起こらなかっとことを表す．

生存曲線の描写：Kaplan-Meier法



ここで注意点は，365日のところで，矩形の上側を見るのか下側を見るのかという点である．Kaplan-Meier曲線の矩形は上側ではなく下側での生存率を見なければならない．

JMPの生命表で生存割合を探る

組み合わせ

time	生存率	故障率	生存標準誤差	故障数	打ち切り数	リスク集合の大きさ
0.00	1.0000	0.0000	0.0000	0	0	228
5.00	0.9956	0.0044	0.0044	1	0	228
11.00	0.9825	0.0175	0.0087	3	0	227
12.00	0.9781	0.0219	0.0097	1	0	224
13.00	0.9693	0.0307	0.0114	2	0	223
⋮	⋮	⋮	⋮	⋮	⋮	⋮
305.00	0.5129	0.4871	0.0350	1	0	87
306.00	0.5070	0.4930	0.0351	1	0	86
310.00	0.4950	0.5050	0.0352	2	0	85
315.00	0.4950	0.5050	0.0352	0	1	83
320.00	0.4890	0.5110	0.0353	1	0	82
⋮	⋮	⋮	⋮	⋮	⋮	⋮
361.00	0.4278	0.5722	0.0358	1	0	70
363.00	0.4154	0.5846	0.0358	2	0	69
364.00	0.4092	0.5908	0.0358	1	1	67
371.00	0.3966	0.6034	0.0358	2	0	65
376.00	0.3966	0.6034	0.0358	0	1	63
⋮	⋮	⋮	⋮	⋮	⋮	⋮
707.00	0.1246	0.8754	0.0290	1	0	15
728.00	0.1157	0.8843	0.0283	1	0	14
731.00	0.1068	0.8932	0.0275	1	0	13
735.00	0.0979	0.9021	0.0266	1	0	12

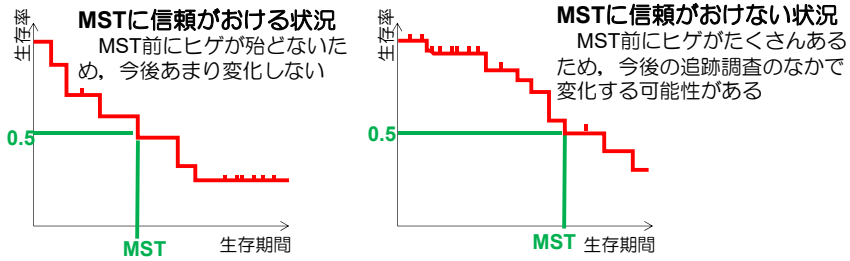
中央生存期間 = **310日**

1年生存割合(365.25日) = **40.92%**

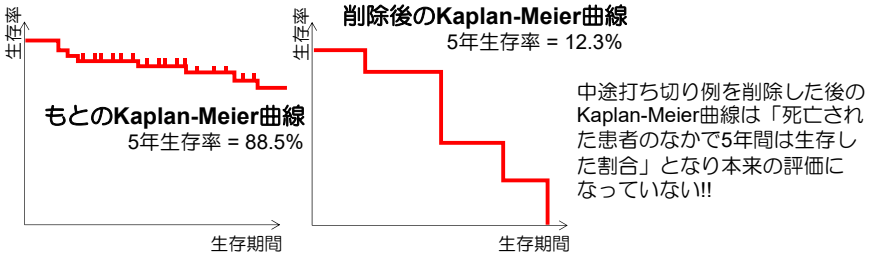
2年生存割合 (365.25×2=730.5日) = **11.57%**

Kaplan-Meier曲線の意味

■ 中途打ち切りを表すヒゲには意味がある

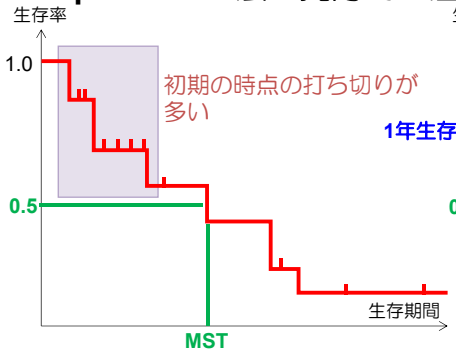


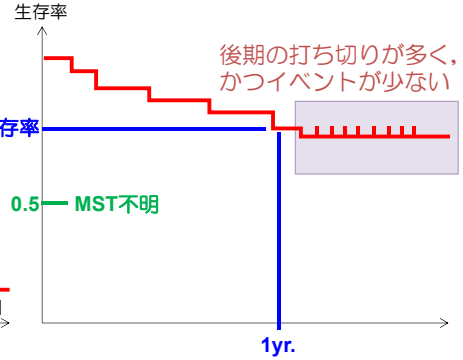
■ 中途打ち切り例を削除してはならない



中途打ち切り例を削除した後のKaplan-Meier曲線は「死亡された患者のなかで5年間は生存した割合」となり本来の評価になっていない!!

Kaplan-Meier法の見方での注意点





初期の時点で打ち切りが多いということは，きっちりとしたフォローアップがなされていない，あるいは，研究実施体制に問題があると判断される可能性がある．

また，この場合のMSTは不完全である．

後期の時点で打ち切りが多く，かつフォローアップ期間中にイベントが少ない理由：

① 対象疾患の予後が良好な疾患の場合

②フォローアップ期間が十分でない場合がある．

この場合には，MSTは計算できないため，年次生存率のみを計算

少し脱線：エンドポイントについて

がん臨床研究での真のエンドポイントの一つは，全生存期間である．一方で，全生存期間には，

- 予後良好な場合，打ち切り(censoring)が多くなり，不完全データになりやすい．
- 長期のフォローアップ期間が必要になる．
- 後続ラインの影響を受ける．とくに，治療法がクロスオーバーする場合は深刻である．

のような問題があることから，代替エンドポイント(サロゲートエンドポイント)が用いられる．その代表的なものの一つが無増悪生存期間(PFS)である．

全生存期間は治療効果を「常に」評価し得るものか？

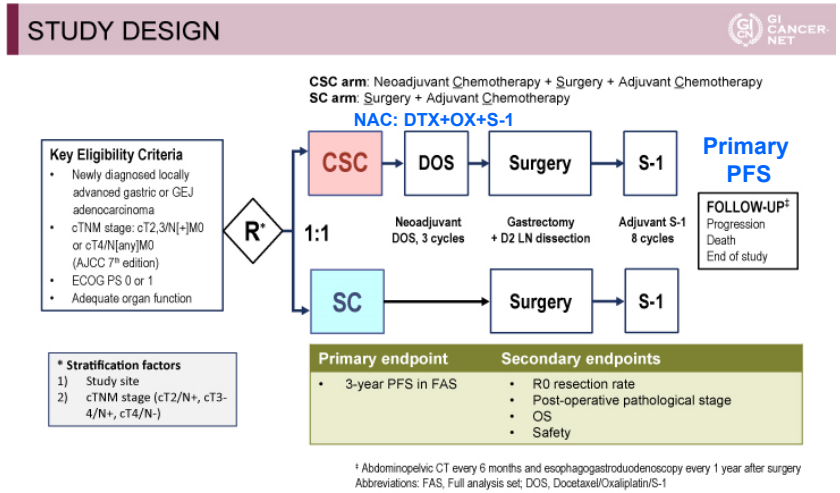
JCOGにおけるエンドポイントの定義

エンドポイント	イベント (いずれか早いもの)			打ち切り日
全生存期間 Overall Survival (OS)	あらゆる死亡	—	—	最終生存確認日
無増悪生存期間 Progression-Free Survival (PFS)	あらゆる死亡	増悪／再発	—	最終無増悪確認日
無発生生存期間 Relapse-Free survival (RFS)	あらゆる死亡	再発	—	最終生存確認日 (最終無再発確認日)
無病生存期間 Disease-Free survival (DFS)	あらゆる死亡	再発	2次がん	最終生存確認日 (最終無再発確認日)
治療生存期間(治療完了なし) Time to Treatment Failure (TTF)	あらゆる死亡	治療中止	—	最終治療継続確認日
治療生存期間(治療完了あり) Time to Treatment Failure (TTF)	あらゆる死亡	治療中止	プロトコル治療完了後の増悪／再発	最終治療継続確認日 (プロトコル治療中) 最終無増悪確認日 (プロトコル治療後)

- 術後補助化学療法において，PFSを利用することは多いように思われる．
- 術前補助化学療法，手術については，DFSを利用することが多いように思われる．
- 術後補助化学療法の継続性を評価することを意図して，TTFを副次的に評価することも増えてきている．

この研究のエンドポイントをどう思いますか？

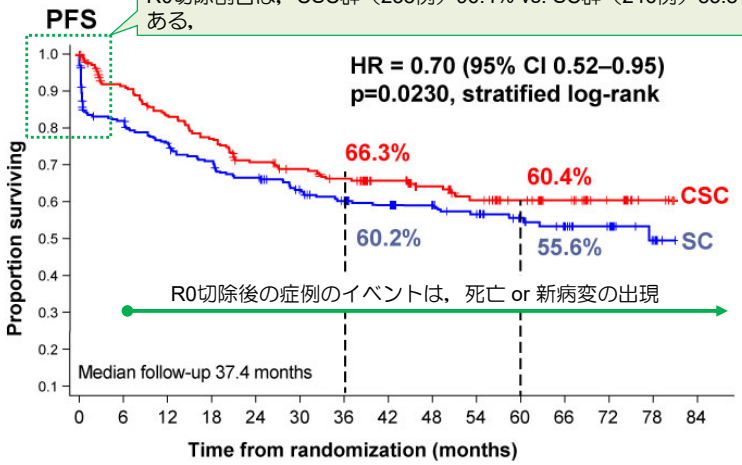
切除可能進行胃癌に対する、手術＋術後S-1療法vs. Docetaxel（DTX）＋Oxaliplatin（OX）＋S-1併用の術前補助化学療法（DOS療法）＋手術＋術後S-1療法：PRODIGY試験



Yoon-Koo Kang, et al. : A Phase III open label randomized study of neoadjuvant chemotherapy with docetaxel, oxaliplatin and S-1 (DOS) followed by surgery and adjuvant S-1, vs surgery and adjuvant S-1 for resectable advanced gastric cancer (PRODIGY Study), ESMO Congress 2019

この研究のエンドポイントをどう思いますか？

R0切除割合は，CSC群（238例）96.4% vs. SC群（246例）85.8%である，



ちなみに，OSはCSC群のイベントが少なく(27.3%)，HR＝0.84（95% CI: 0.60-1.19）,p＝0.3383

本試験のPrimary endpointはPFSでよかったのでしょうか？ 良くないならば，何が良かったのでしょうか ？

生存期間を取り扱ううえで重要な測度(関数)

■ 生存時間関数

個体が少なくとも時点 t で生存する確率

時点tまで生存した割合．
年次生存割合などは，生存時間関数を用いて計算する

■ ハザード関数

個体が時点tまで生存しているという仮定のもとで微小区間

$t \le T < t + \Delta t$で死亡する確率

時点tにおける死亡リスク．
症例数設計などは，ハザード関数を用いて計算する

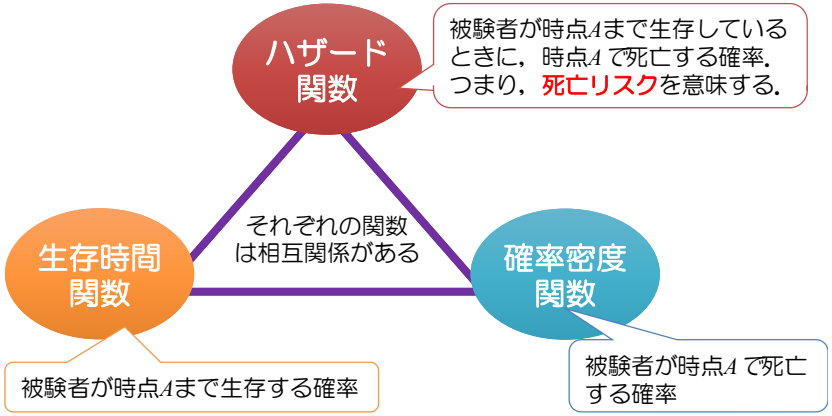
■ 確率密度関数 (probability density function)

個体が微小区間 $t \le T < t + \Delta t$内で死亡する確率

■ 累積ハザード関数 (cumulative hazard function)

ハザード関数の累積値

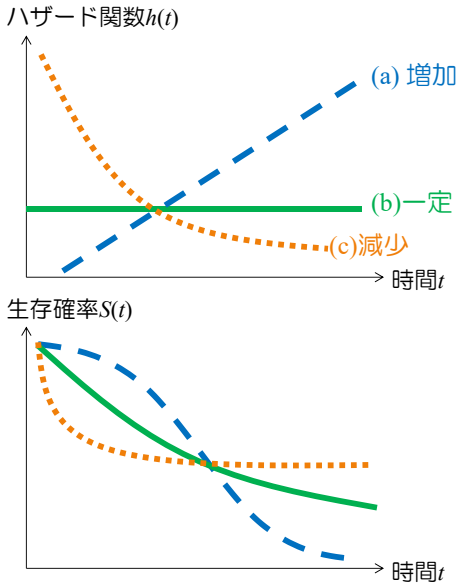
生存時間解析において重要な関数



確率密度関数，確率密度関数，生存時間関数は相互関係があり，どれか1個が決まれば，すべてが決定できる．

ログランク検定ではハザードを評価しているのに，生存時間曲線を描写してグラフを描写している．これは，3種類の関数は相互関係があるため．

ハザードの意味



(a) の状況
登録時点では死亡(あるいはイベント)のリスクは低いものの時間経過とともに増大することを意味する．

(b) の状況
全時間上でハザード(すなわちリスク)が一定であることを意味する．

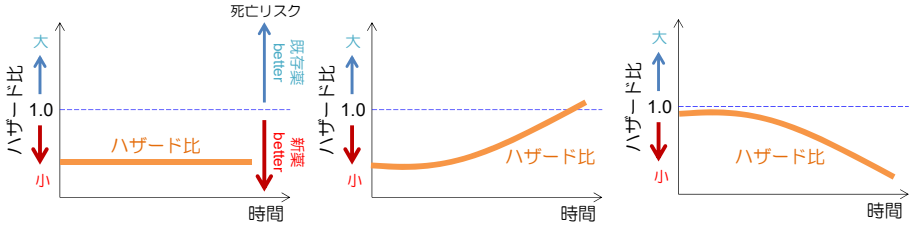
(c) の状況
登録時点では死亡(あるいはイベント)のリスクは高いものの時間経過とともに減少することを意味する．一般に，進行がんに対して見られる．

生存曲線の比較において重要な指標：ハザード比

ハザード(危険度) $\lambda(t)$とは「時間tまで患者が生存しているという条件のもとで，時間tで死亡する確率」を表す．つまり，時間tでの死亡リスクである．

ハザード比(相対危険度)とはハザードの比である．例えば，既存薬のハザードに対する新薬のハザードでのハザード比は

$$\text{ハザード比 } HR(t) = \frac{\text{新薬でのハザード } \lambda_1(t)}{\text{既存薬でのハザード } \lambda_0(t)}$$

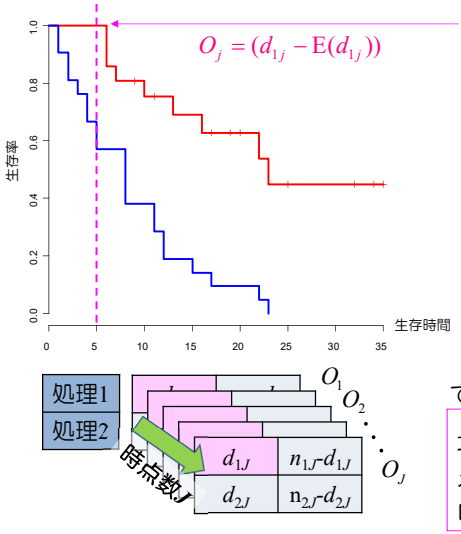


ハザード比が生存時間に依らず一定．これは，新薬のほうが死亡リスクが低い．この状況が一般的に想定される(比例ハザード性の仮定)．

ハザード比が生存時間とともに上昇，これは，新薬での死亡リスクが生存時間の短い患者では低いものの，その有効性は長期生存症例になるにつれて薄れる．

ハザード比が生存時間とともに下降，これは，新薬での死亡リスクが生存時間の短い患者では既存薬と大差がないものの，その有効性は長期生存症例になるにつれて強くなる．

ログランク検定とは何をしているか？



ある時点t_jにおける死亡・生存に対するクロス集計表

	死亡	生存	合計
処理１	d_{1j}	$n_{1j}-d_{1j}$	n_{1j}
処理２	d_{2j}	$n_{2j}-d_{2j}$	n_{2j}
合計	d_j	n_j-d_j	n_j

もし，処理1と処理2で死亡数が同じであれば(つまり処理効果がなければ)，■の期待度数は，カイ2乗検定と同じ理屈から

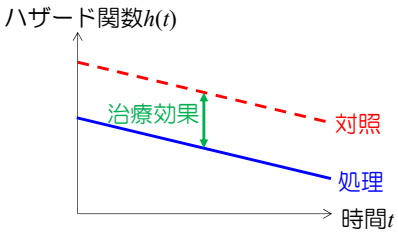
$$\mathrm{E}(d_{1j}) = \frac{d_j n_{1j}}{n_j}$$

である．期待度数からの乖離

上記の2×2クロス集計表の■部分をすべての時点において計算し，その総和をとる$O_1 + O_2 + \cdots + O_J$ことで構成されたのが，ログランク検定である．

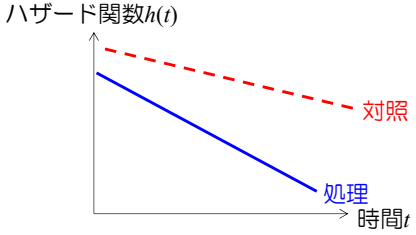
ログランク検定における仮定：比例ハザード性

○ 比例ハザード性



時間tに関係なく治療効果の差が一致である.

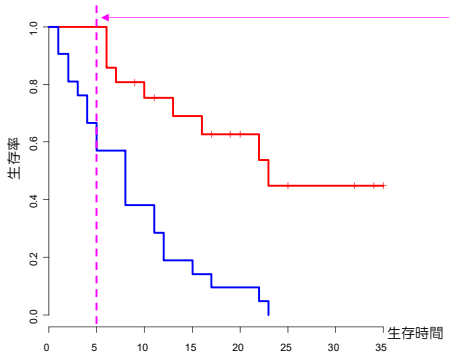
× 比例ハザード性



時間tに関係なく治療効果の差が一致でない.

比例ハザード性の仮定は，ログランク検定だけでなく比例ハザードモデルにおいてもおかれている.

もうひとつの検定方法：一般化Wilcoxon検定



ログランク検定 $(d_{1j} - \mathrm{E}(d_{1j}))$

一般化Wilcoxon検定

$\hat{S}(t_j) \times (d_{1j} - \mathrm{E}(d_{1j}))$

一般化Wlcoxon検定では，ログランク検定に対して，生存率による重み付けを行っている(生存率の推定がKaplan-Meier法の場合にはPeto-Petoの一般化Wilcoxon検定，Prentice法の場合にはPrenticeの一般化Wilcoxon検定と呼ばれる).

時間tが小さい場合には，$\hat{S}(t_j)$ が1に近い(重みが大きい)ことから，一般化Wilcoxon検定では，早期の2群間の差に敏感である.

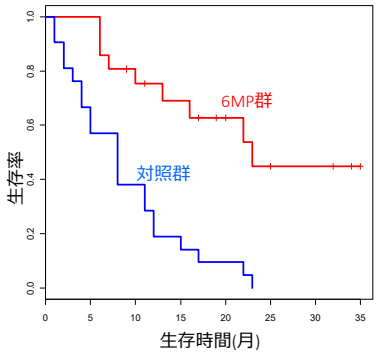
一般化Wilcoxon検定の用途

2標本の比較の例示(1)

小児急性白血病データ
(Freireich *et al.*, 1963)

6MP投与群	6	6	6	6+	7	9+	10	10+
	11+	13	16	17+	19+	20+	22	23
	25+	32+	32+	34+	35+			
対照群	1	1	2	2	3	4	4	5
	5	8	8	8	8	11	11	12
	12	15	17	22	23			

+は中途打ち切り，単位は月



諸種の検定におけるP値

ログランク検定 p<0.0001
一般化Wilcoxon検定 p=0.0002

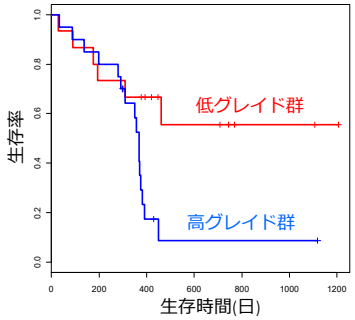
いずれの検定法でも6MP群と対照群のあいだで有意であることがわかる.

2標本の比較の例示(2)

卵巣がん患者のグレイド毎での病気進行期間(Fleming *et al.*, 1980)

低グレイド群	28	89	175	195	309	377+	393+	421+
	447+	462	709+	744+	770+	1106+	1206+	
高グレイド群	34	88	137	199	280	291	299+	300+
	309	351	358	369	369	370	375	382
	392	429+	451	1119+				

+は中途打ち切り，単位は日

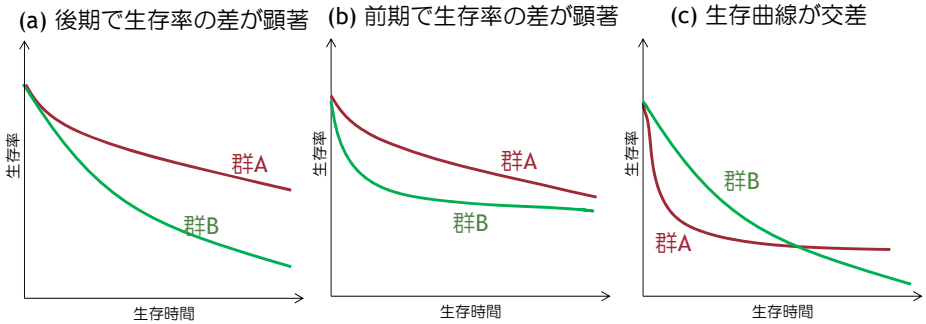


諸種の検定におけるP値

ログランク検定 p=0.0183
Gehanの一般化Wilcoxon検定 p=0.0995

比例性は満たされていないと指摘されているデータではあったものの，ログランク検定のみに有意差が認められた．初期の生存時間で差が認められなかったことが寄与していると考えられる.

手法による違い(大橋・浜田,1995)



(a)の状況では，ログランク検定のほうが一般化Wilcoxon検定よりも検出力が強い．

(b)の状況では，一般化Wilcoxon検定のほうがログランク検定よりも検出力が高い．

(c)の状況では，いずれの方法も検出力が低い．

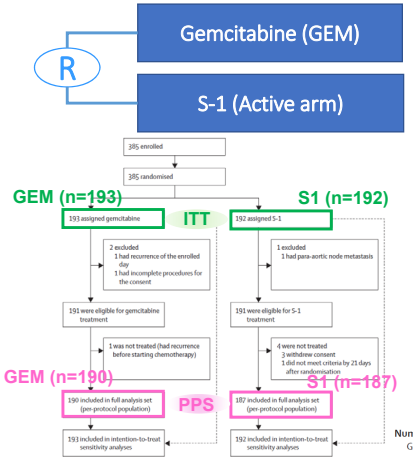
検出力とは，治療効果があるときに「有意である」と正しく判断できる確率を意味する．

少し脱線：ITT(Intent-To-Treat)の原則とPPS(Per Protocol Set)

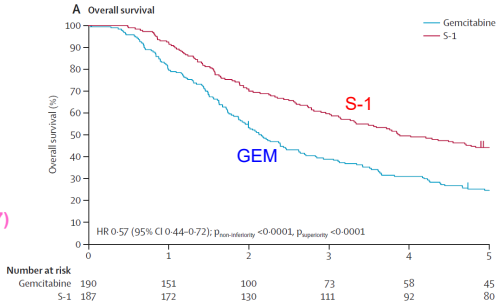
■ JASPAC 01 Study (Uesaka et al., Lancet, 388, 248-257, 2016)

治癒切除された浸潤性膵管癌患者に対する GEM vs. S-1の無作為化比較非劣勢試験

Primary endpoint ： PPSにおける全生存期間(OS)

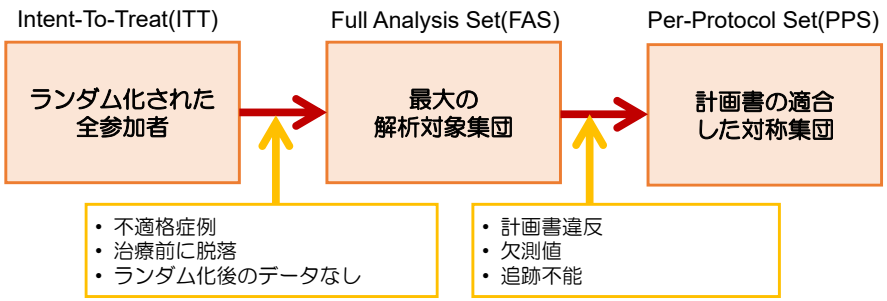


GEM群の3年生存割合を36%と仮定し，S-1群のGEM群に対するハザード比を0.87と見込み，片側α=0.025，検出力80%（非劣性マージン1.25）で240イベントが必要である．登録期間3年・観察期間2年とした場合に，必要症例数は360例と計算されている．



無作為化比較試験では，ITT症例で解析することが一般的では？

少し脱線：ITT(Intent-To-Treat)の原則とPPS(Per Protocol Set)



■ 選択バイアスの観点から，原則的には，ITT症例(or FAS)での解析が行われる(ICH-E9ガイドラインにも記載)．

■ 純粋に治療法(あるいは薬剤)の評価を行う場合には，PPSを主解析とする場合がある．例えば，安全性が主要評価項目の場合である．

■ また，非劣性試験では，研究の質が悪いと群間差が小さくなることや，プロトコルのアドヒアランスを評価するために，PPSで実施することがある．JASPAC01において，PPSを利用したのはそのため．

共変量調整による生存時間分布の比較

共変量調整による比較の主な狙い

- 応答と共変量の間に何らかの強い関連のあることが既知であるとき，共変量の応答への影響を排除することで個体間のバラツキを縮小させ，比較の精度をあげる．
- 共変量の個々の層で明快な治療差が出なくても，幾つかの層で治療差が同一方向を示せば，治療差の検出感度を高めることができる．

共変量が必要な場面

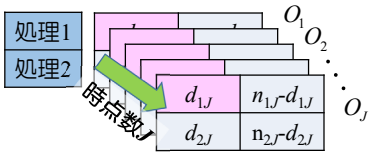
- 層別因子を含む確率化臨床試験
- 共変量の治療群間での偏りの調整(事後層別)

生存時間分布の比較における共変量調整方法

- 既存の順位検定の拡張(層別ログランク検定)
- 比例ハザードモデルに基づく尤度比検定

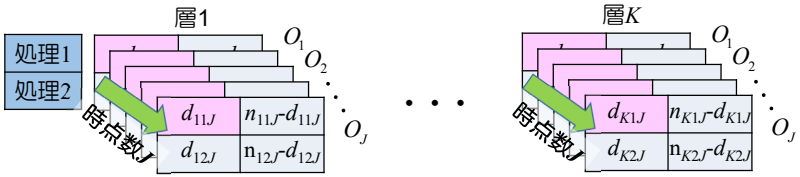
層別ログランク検定

■ ログランク検定



複数の2×2クロス集計表に基づいて，検定しているのがログランク検定．

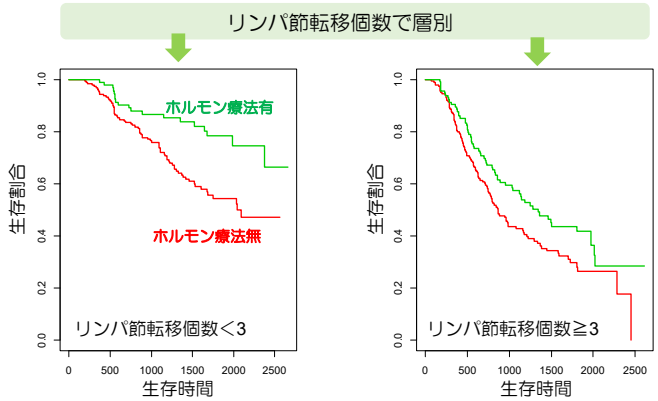
■ 層別ログランク検定



ログランク検定は，時間方向に作成された複数の２×２分割表を平均化したものである．層別ログランク検定，共変量の組み合わせによるそれぞれの層に分割し，そのもとで，時間方向に作成された複数の２×２分割表を平均化することで構成される．

層別ログランク検定の例

ドイツで実施された，ホルモン療法の有無による有効性・安全性を評価するための無作為化比較第III相試験の例

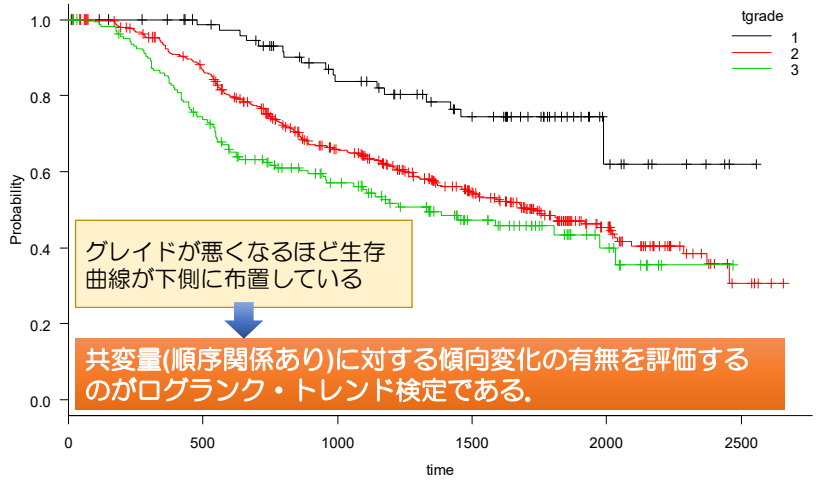


リンパ節転移個数が多いほど予後不良である．もし，群間でリンパ節転移個数に偏りがあると，治療効果なのか，リンパ節転移個数が影響したのかわからない． ➡ 層別ログランク検定

ログランク検定	層別ログランク検定
p-value=0.003	p-value<0.001
HR 0.695 [0.544, 0.888]	Adj. HR 0.650 [0.508, 0.831]

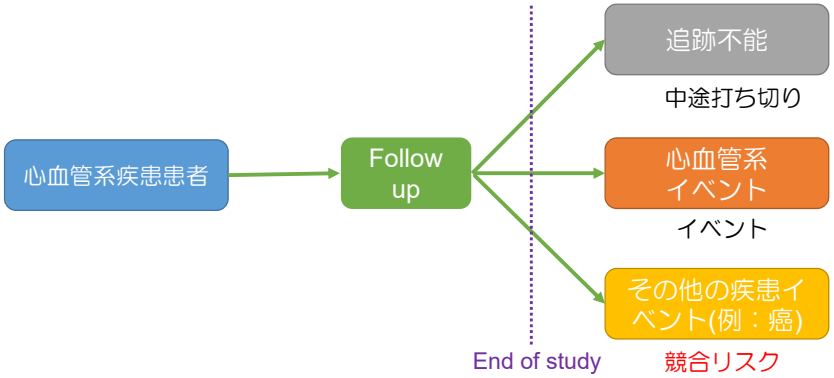
傾向性の検定：ログランク・トレンド検定

ホルモン療法の効果を検討するために，ドイツ乳癌研究グループ(GBSG; German Breast Cancer Study Group)が実施した無作為化比較第III相試験の結果である．
このデータは，腫瘍グレイド(tgrade)が1から3までとられている．
ここでの関心は，主要グレードが高くなるにつれて，生存曲線が下側にいく傾向にあるか否かを検定することにある．



打ち切りに情報がある場合：競合リスク (Competing risk)

競合リスクの例：心血管系の患者に対する治療法の評価

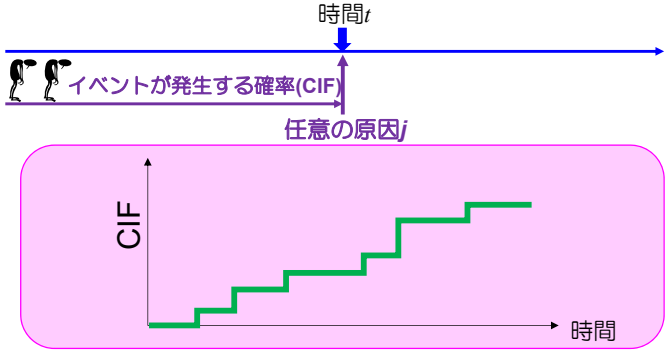


心血管系患者が心血管系イベントの前に，癌などの疾患により死亡した場合，心血管系治療後のイベントまでの期間が不明になる．このとき，癌によるイベント発現との関係を**競合リスク(competing risk)**という．

競合リスクでは生存率ではなく発生率を評価する

累積発生関数(CIF; Cumulative Incident Function)

原因jによるイベントが時間tまでに発現する確率(Kalbfleisch & Prentice, 1980)



通常の生存時間解析では，
生存率：任意の生存期間までに，生存(イベントがない)している割合
を評価するが，競合リスクの場合にはイベントにラベリングが行われているので，
累積発現率：任意の生存期間までに，任意のイベントが発現する割合
を評価する．

競合リスクが存在する場合の生存時間解析：累積発生関数

通常の生存時間解析
Kaplan-Meier法
時間tまで生存している割合を評価する
(例：1年生存割合，中央生存期間)

→

競合リスクがある場合
累積発生関数(CIF)
時間tまでに関心のあるイベントが発生した割合を評価する(例：1年発生割合)

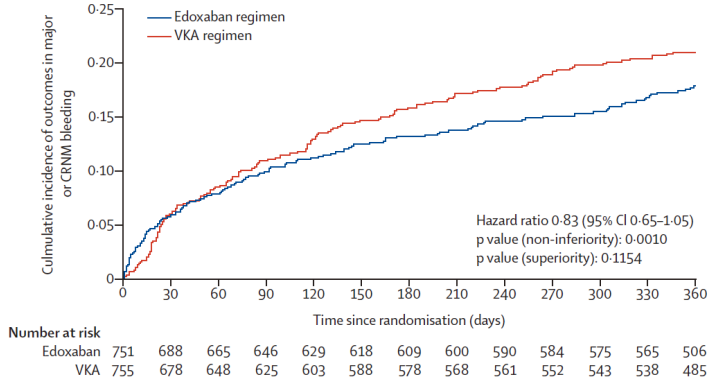
			Kaplan-Meier 競合リスクを打ち切り		CIF
時間 t	イベント /中途打ち切り	リスク集合 d_{in}	無発症率 (event free survival)	イベント発現率	イベント発現率 I_1
100	イベント	9	1.00 × (8/9) = 0.89	0.00+1.00×(1/9) = 0.11	0.00+1.00×(1/9) = 0.11
200	中途打ち切り	8	0.89 × (8/8) = 0.89	0.11+0.89×(0/8) = 0.11	0.11+0.89×(0/8) = 0.11
300	イベント	7	0.89 × (6/7) = 0.76	0.11+0.89×(1/7) = 0.24	0.11+0.89×(1/7) = 0.24
400	競合リスク	6	0.76 × (6/6) = 0.76	0.24+0.76×(0/6) = 0.24	0.24+0.76×(0/6) = 0.24
500	イベント	5	0.76 × (4/5) = 0.61	0.24+0.76×(1/5) = 0.39	0.24+0.63×(1/5) = 0.37
600	競合リスク	4	0.61 × (4/4) = 0.61	0.39+0.61×(0/4) = 0.39	0.37+0.50×(0/4) = 0.37
700	競合リスク	3	0.61 × (3/3) = 0.61	0.39+0.61×(0/3) = 0.39	0.37+0.38×(0/3) = 0.37
800	競合リスク	2	0.61 × (2/2) = 0.61	0.39+0.61×(0/2) = 0.39	0.37+0.25×(0/2) = 0.37
900	中途打ち切り	1	0.61 × (1/1) = 0.61	0.39+0.61×(0/1) = 0.39	0.37+0.12×(0/1) = 0.37

CIFを計算する場合，直感的には，1-KMEと考えられるが，そうすると，Over estimate(イベント発現率を高く見積もる)ことになる．

累積発生関数の適用例

安定冠動脈疾患または急性冠症候群に対するPCIを受けた心房細動（AF）患者1506例を対象に，Edoxeban(エドキサバン) + P2Y12阻害薬 vs VKA(ビタミンK拮抗薬)+P2Y12阻害薬+アスピリンを比較した国際共同第IIIb試験(18カ国，186施設)

主要エンドポイント：12カ月間での大出血または重大な非大出血(CRNM)
→ したがって，それ以外のイベントは**競合リスク**になり得る．



Vranckx et al.(2019). Edoxaban-based versus vitamin K antagonist-based antithrombotic regimen after successful coronary stenting in patients with atrial fibrillation (ENTRUST-AF PCI): a randomised, open-label, phase 3b trial, The Lancet, 394, 12-18.

競合リスクにおける検定：Gray検定

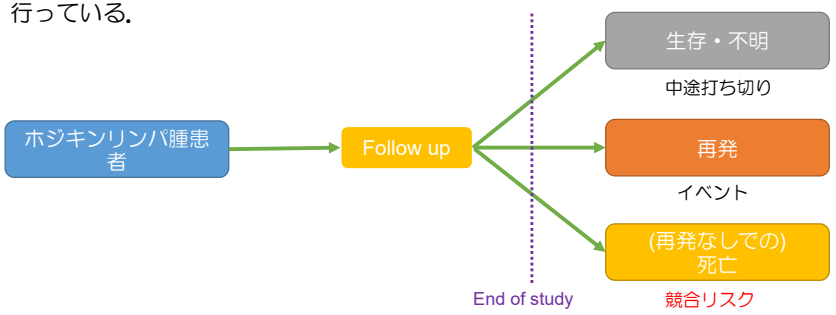
通常の生存時間解析
ログランク検定
帰無仮説H_0「ハザード比が等しい」に対する検定

→

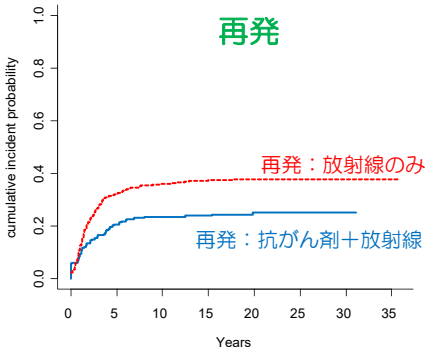
競合リスクがある場合
Gray検定
帰無仮説H_0「累積発生関数が等しい」に対する検定

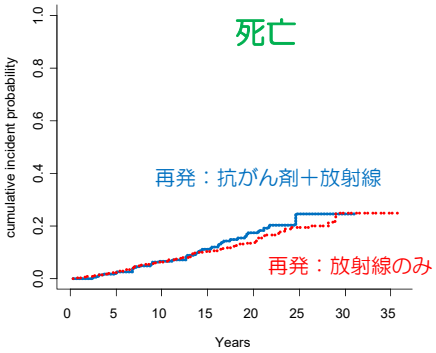
例示：ホジキンリンパ種データ(Pintillie, 2006)

1968年から1986年のあいだにPrincess Margaret病院で調査された，初期(Stage I/II)のホジキンリンパ腫の患者(n=865)のデータである．Pinttillie(2006)は，この調査データを用いて「放射線療法のみ」「放射線療法＋化学療法」の有効性の検討を行っている．



例示におけるGray検定の結果





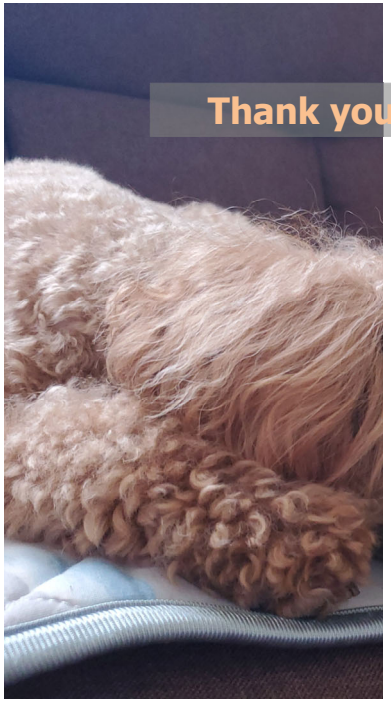
検定統計量 $:\chi_0^2 = 12.685$

→ p値=0.0003

検定統計量 $:\chi_0^2 = 0.729$

→ p値=0.393

すなわち，「抗がん剤＋放射線」併用療法の「放射線のみ」に対する有効性は，再発の抑制に認められたが，再発がない場合の死亡には認められない．



Thank you for your kind attention