# 使いこなそう生物統計！
# ～検定・交絡調整を中心に～

研究支援センター 生物統計部
/東病院臨床研究支援部門
若林 将史

2019. 12.14 (土)

# 試験の結果を正しく解釈したい！

肺がん内科グループ
JCOG0301

**切除不能III期 非小細胞肺癌**
**71才以上**

↓

**ランダム割付**

**（標準治療）** 100例
放射線単独
(RT)

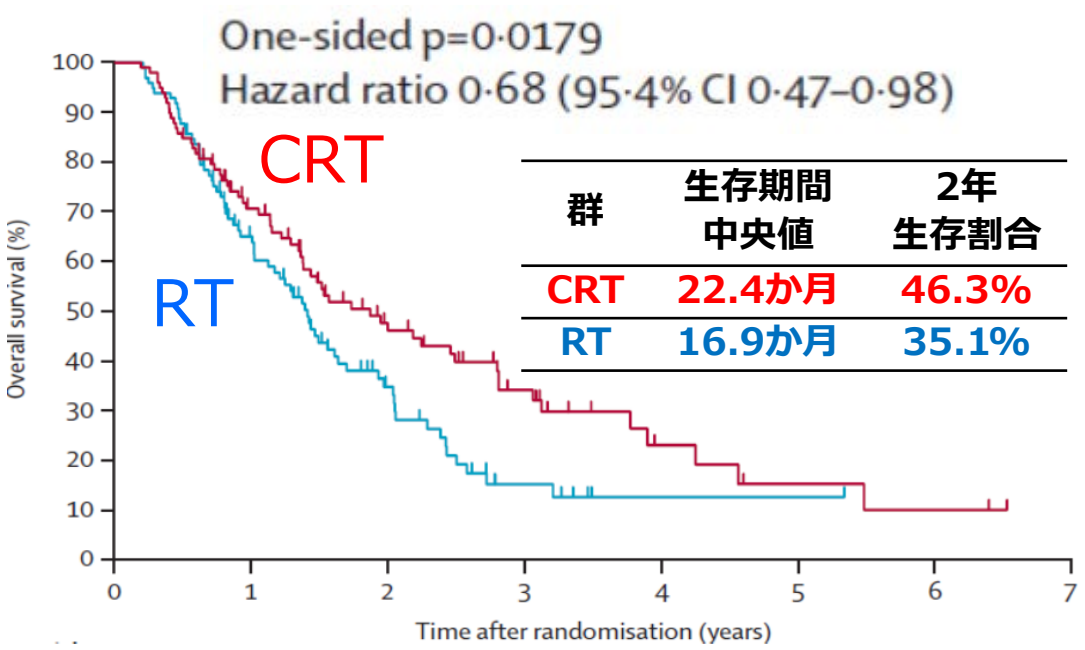
RT施行例 98例
RT完遂 93例
RT非完遂 5例

未治療例 2例

**（新治療）** 100例
放射線+化学療法
(CRT)

CRT施行例 96例
CRT完遂 88例
CRT非完遂 8例

未治療例 4例




| 群 | 生存期間 中央値 | 2年 生存割合 |
|---|---|---|
| CRT | 22.4か月 | 46.3% |
| RT | 16.9か月 | 35.1% |

**結論：CRTはRTと比較して臨床的に意味のあるベネフィットがあり、CRTはこの対象に対して考慮されるべき治療法である**

何故このような結論になるの？

Atagi *et al.* (2012) Lancet Oncology 13(7): 671-8.

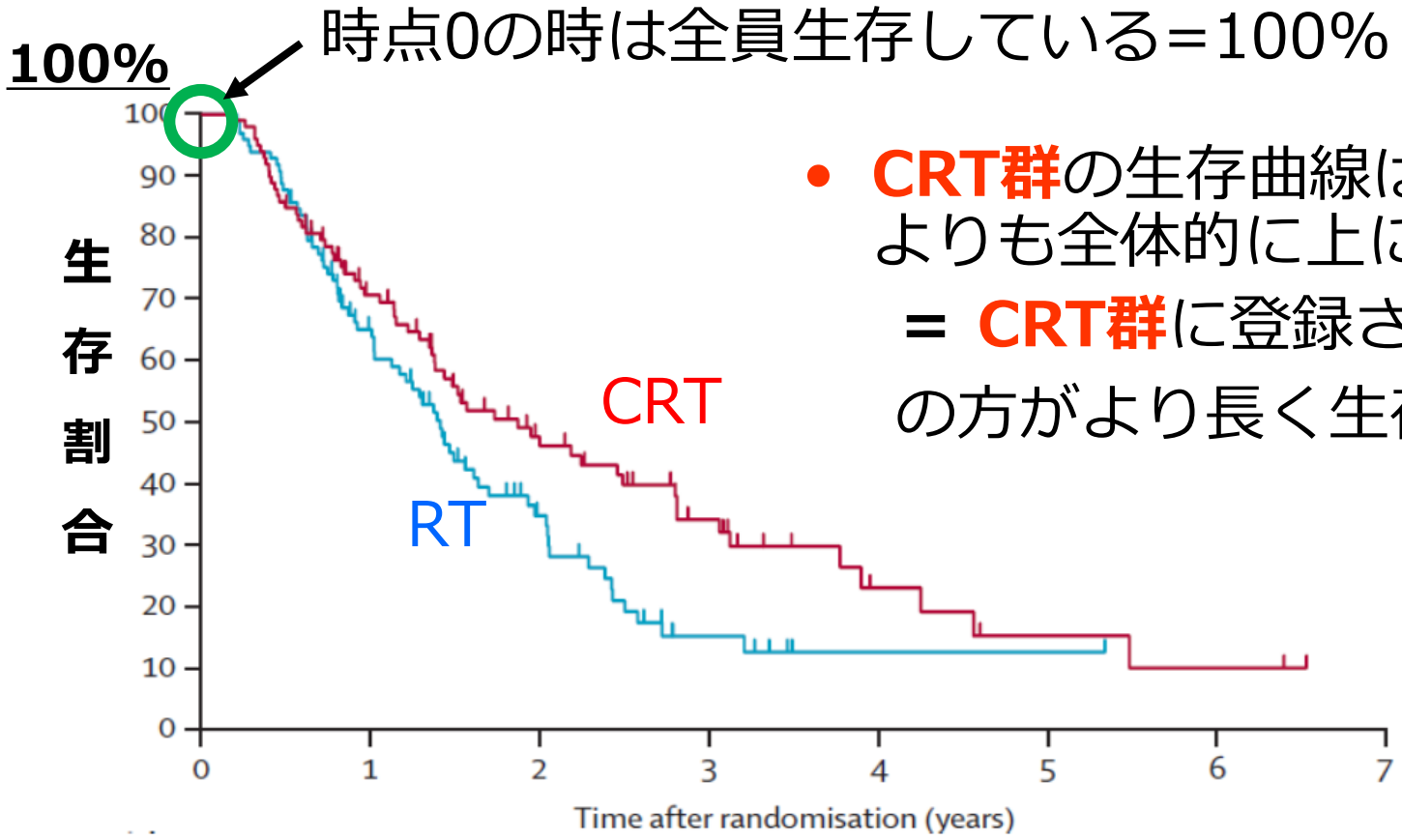
# 講義のTopic：臨床研究に関する結果の解釈

- **ランダム化**はなぜ必要か？
  - **交絡**と**ランダム化**
- **交絡**の調整方法
  - 試験計画段階と解析段階での調整方法
- 結果の検証方法
  - **仮説検定**の考え方と**p値**の意味
  - **αエラー**、 **βエラー**、 **検出力**の理解
- 治療効果の大きさの見方
  - **ハザード比**の意味
- **ITT(Intention-to treat)解析**とは？

# 導入：生存曲線

# 生存曲線とは

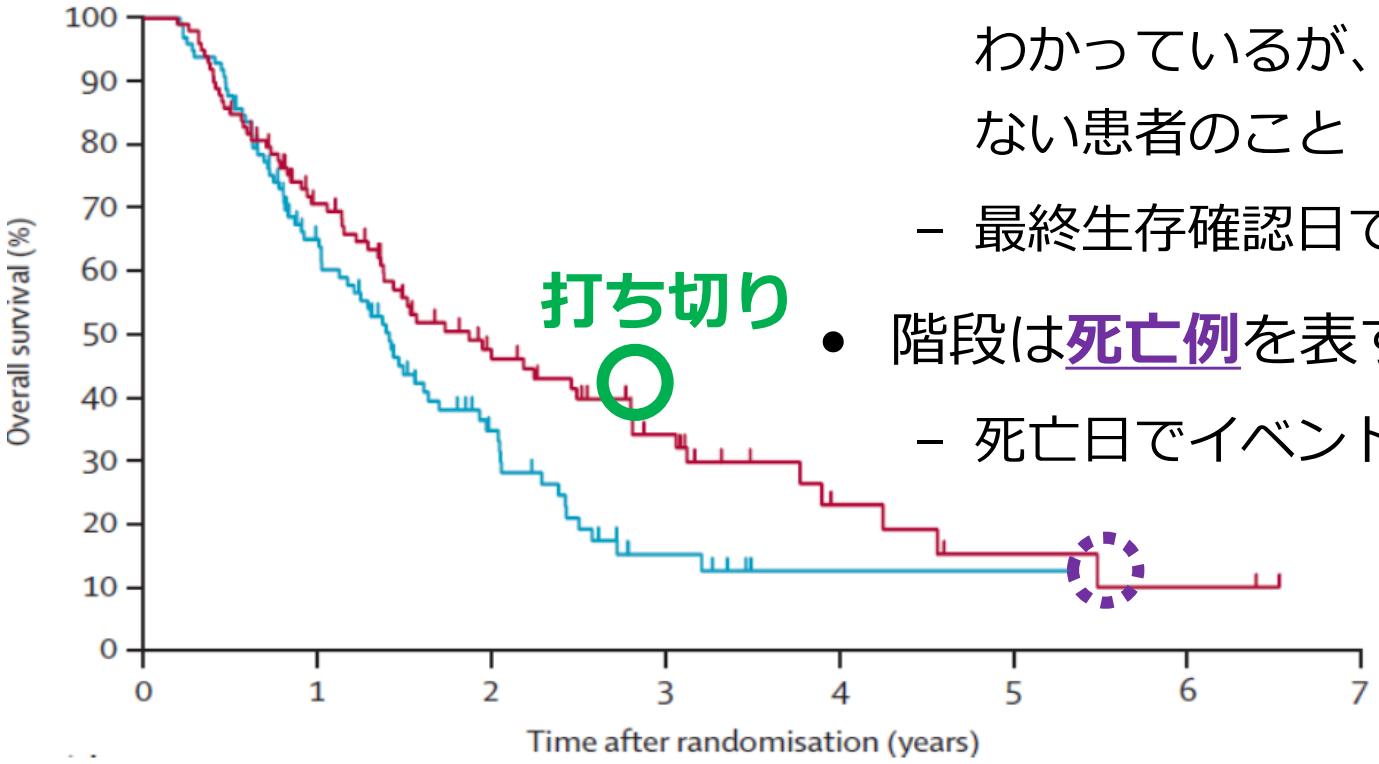
- 縦軸に生存割合、横軸に時間をとり、集団における各時点の生存割合をつないだもの
- 死亡が発生するとその時点で生存割合が減少する




- **CRT群**の生存曲線は**RT群**よりも全体的に上にある
  ＝ **CRT群**に登録された集団の方がより長く生存した

Atagi *et al.* (2012) Lancet Oncology 13(7): 671-8.

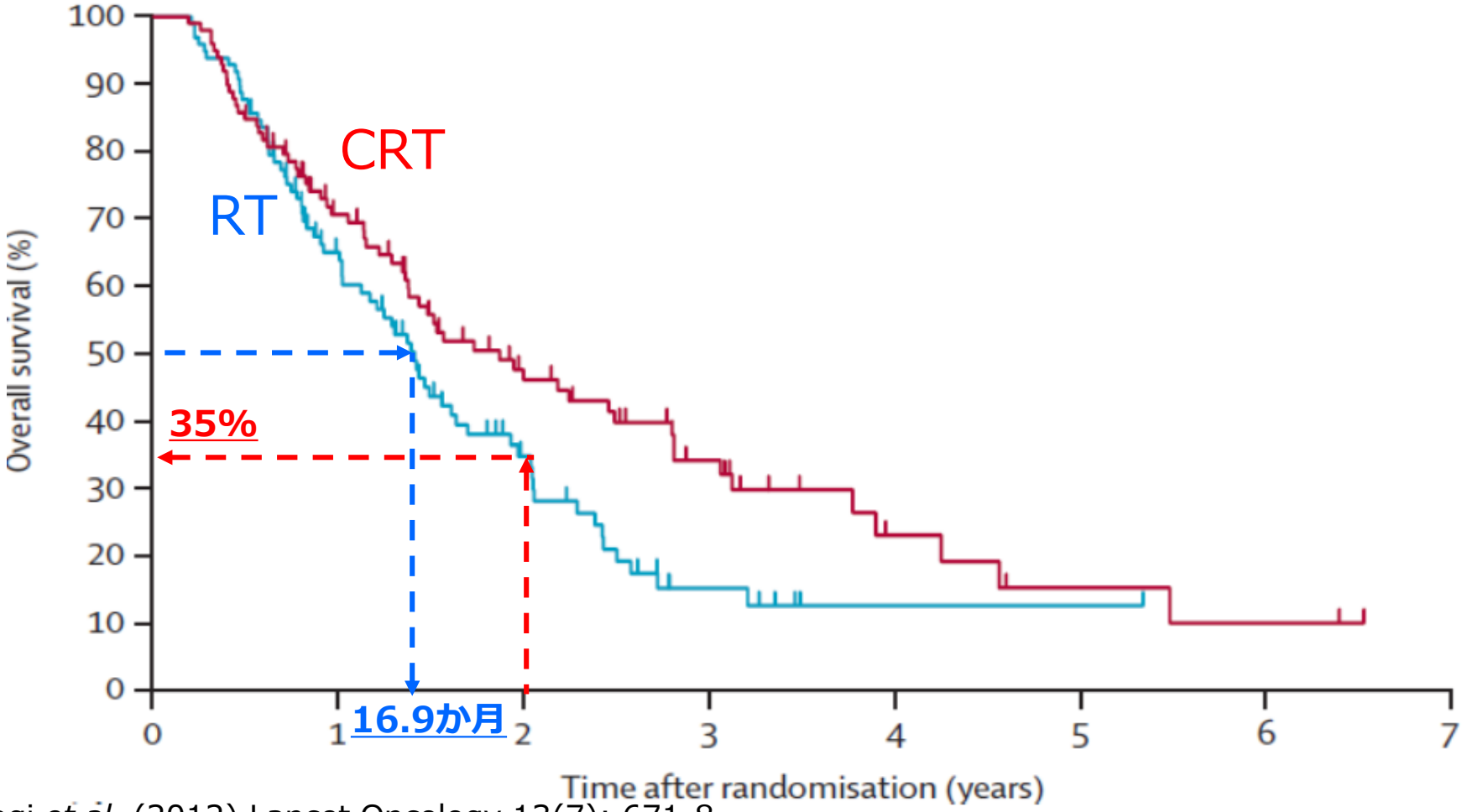
# イベントと打ち切り




- ヒゲの印は**打ち切り例**を表す
  - その時点まで死亡していないことはわかっているが、それ以降の情報がない患者のこと
  - 最終生存確認日で打ち切り
- 階段は**死亡例**を表す
  - 死亡日でイベント

# 生存曲線から得られる要約値

- 生存期間中央値(MST; Median survival time)、時点生存割合
  - RT群の生存期間中央値は**16.9か月**、2年生存割合は**35%**



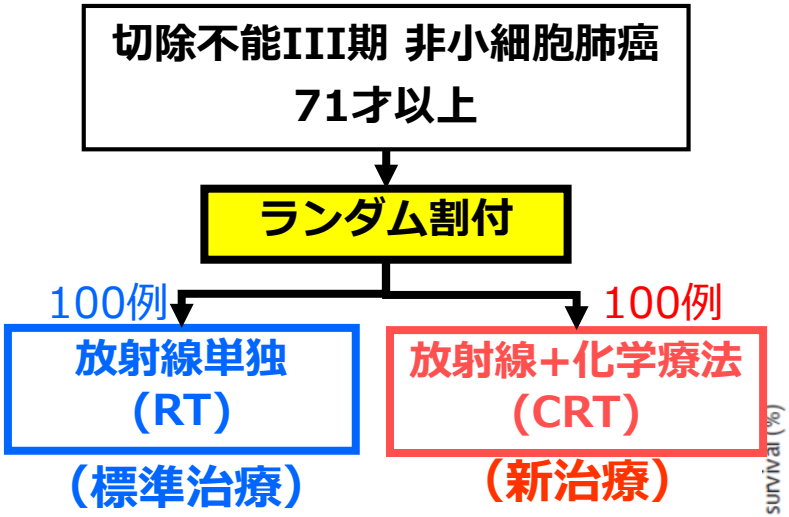

Atagi *et al.* (2012) Lancet Oncology 13(7): 671-8.

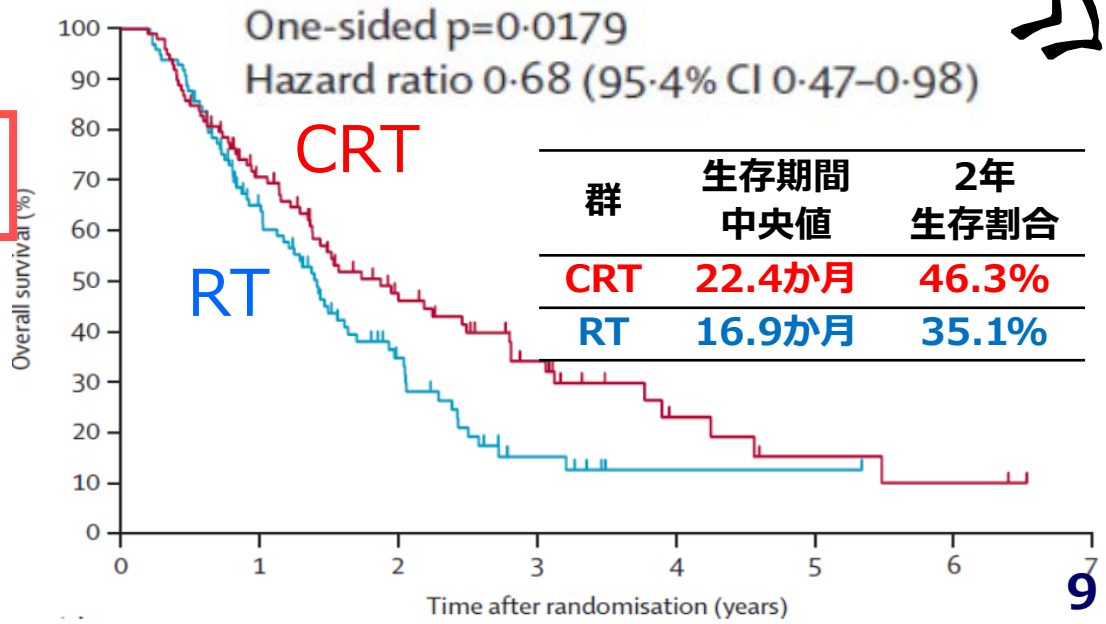
# 交絡とランダム化

# ランダム化？？

ランダム化(ランダム割付)って
何のためにやっているの？
医師や患者が好きな治療をすれば
良いのでは？

肺がん内科グループ
JCOG0301




| 群 | 生存期間 中央値 | 2年 生存割合 |
|---|---|---|
| CRT | 22.4か月 | 46.3% |
| RT | 16.9か月 | 35.1% |

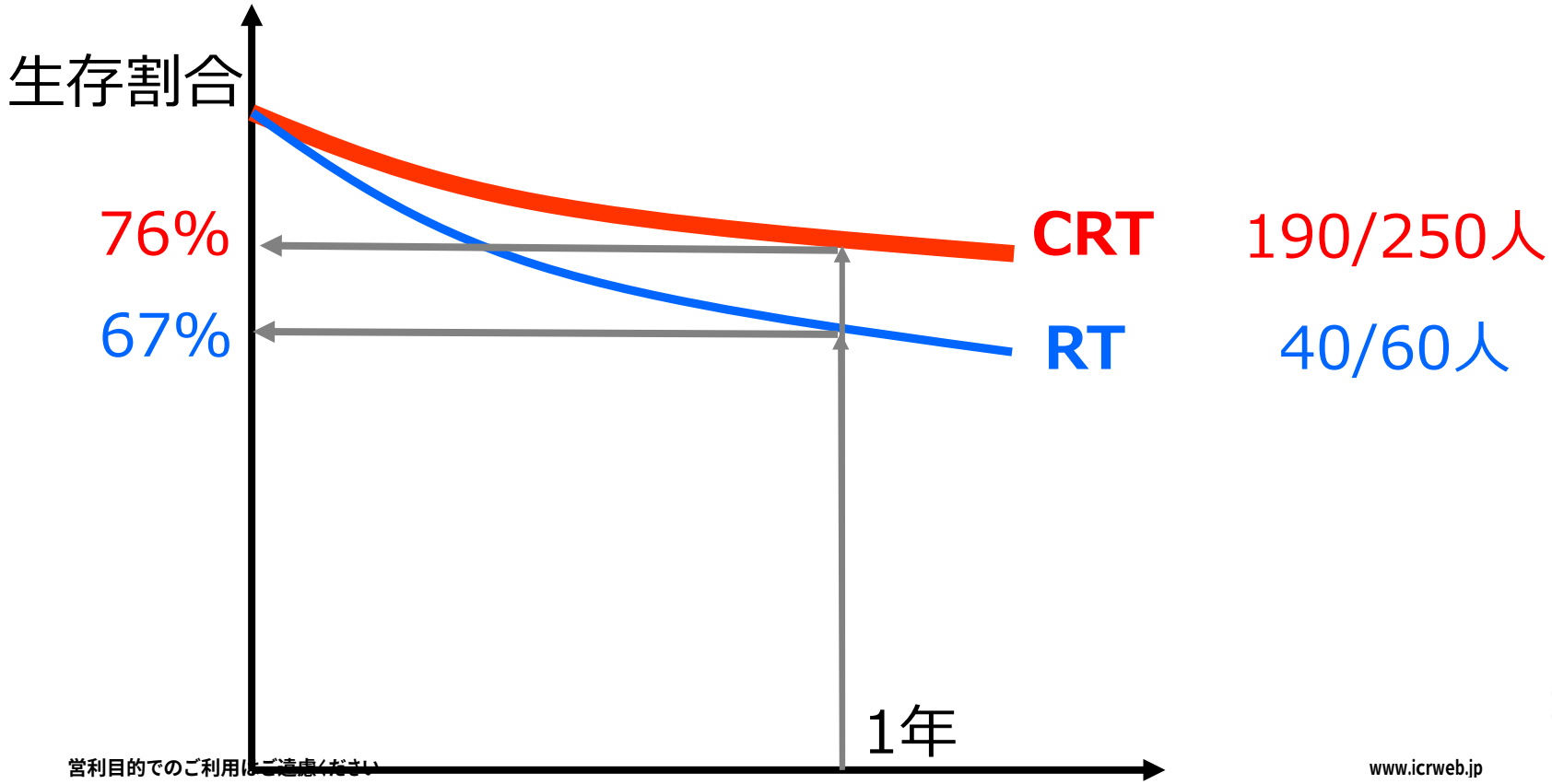


Atagi *et al.* (2012) Lancet Oncology 13(7): 671-8.

注：仮想例

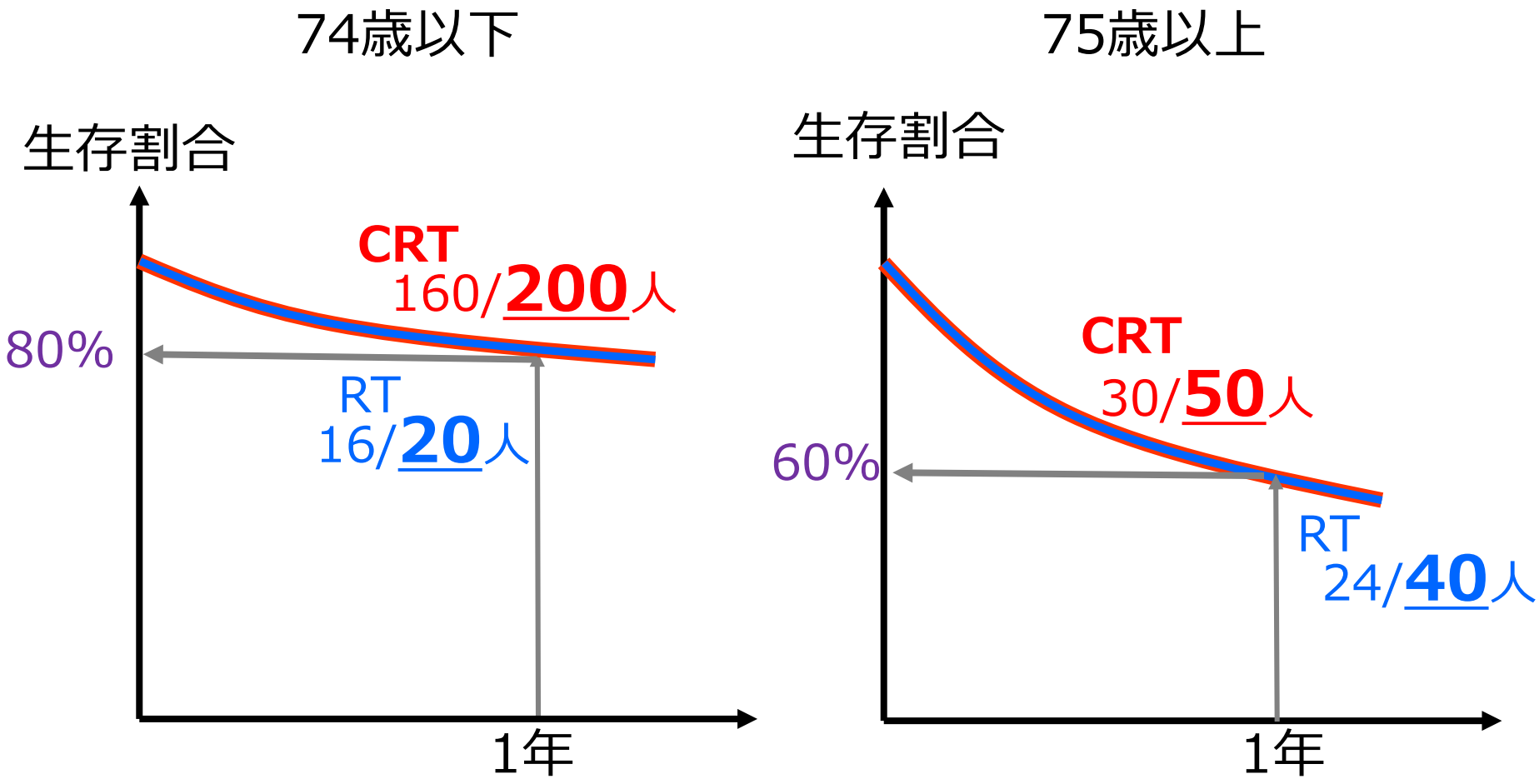
# 国内の学会で見かける発表

- ある病院の適格規準XXを満たす患者を**CRT群**(250例)と**RT群**(60例)に分けてレトロスペクティブに検討した。
- **CRT群**は**RT群**と比較し予後良好であった。
- この対象に**CRT**をすることが推奨される。



注：仮想例

# 年齢で分けた場合の予後




年齢によらずCRTとRTの予後は変わらない

注：仮想例

# 比較したいのはCRTとRTの違いだから

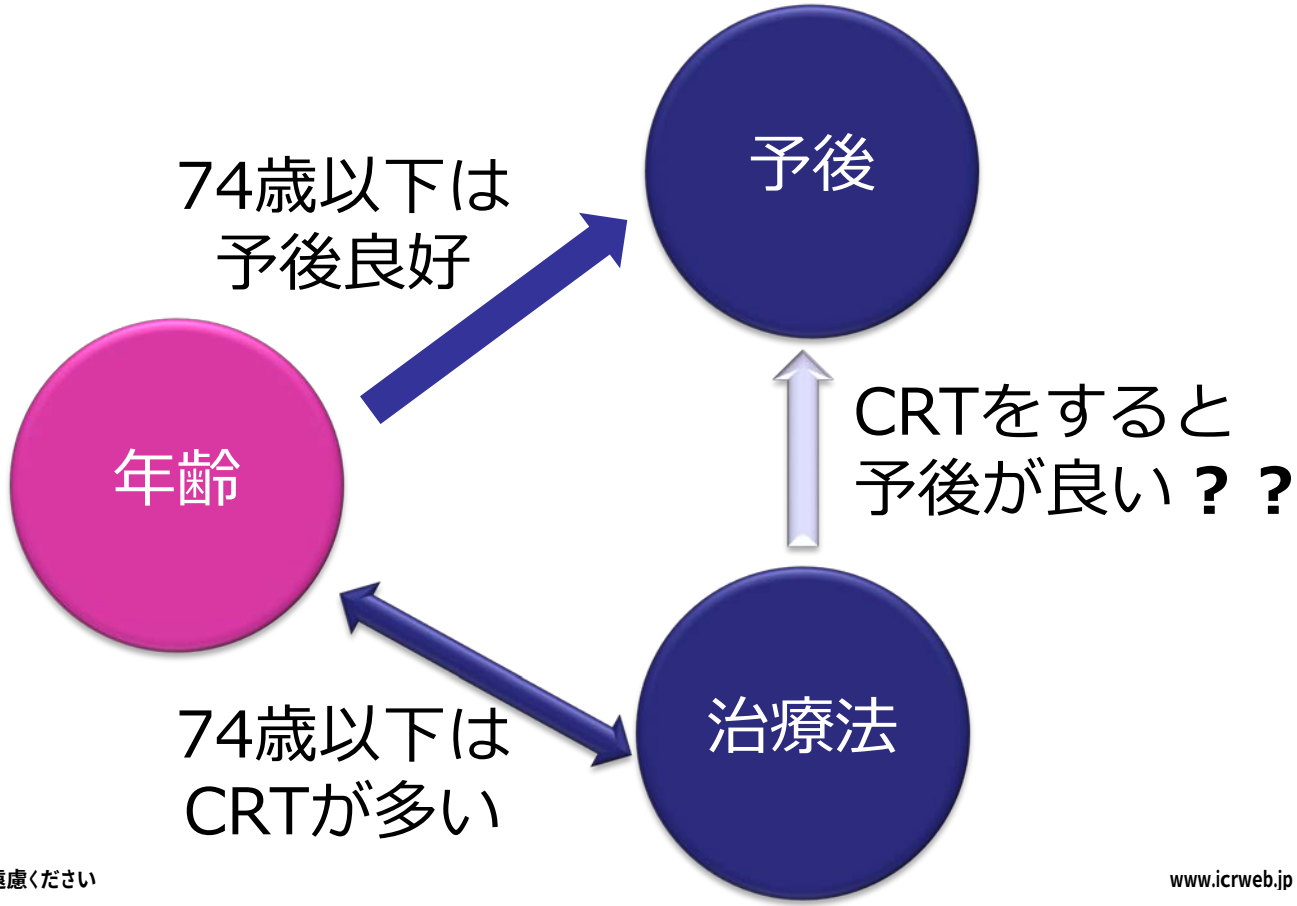
- 治療法以外の予後に影響する要因の条件が同じでなければ、”比較”にならない！！

| 治療法 | 74歳以下 | | 75歳以上 | 合計 |
|---|---|---|---|---|
| CRT | **200人** (80%) | ＞＞ | 50人 | 250人 |
| RT | 20人 (33.3%) | ＜＜ | **40人** | 60人 |

- CRTはRTと比べ**「74歳以下」**の割合が高い
- 年齢によって予後が異なる(74歳以下は予後良)

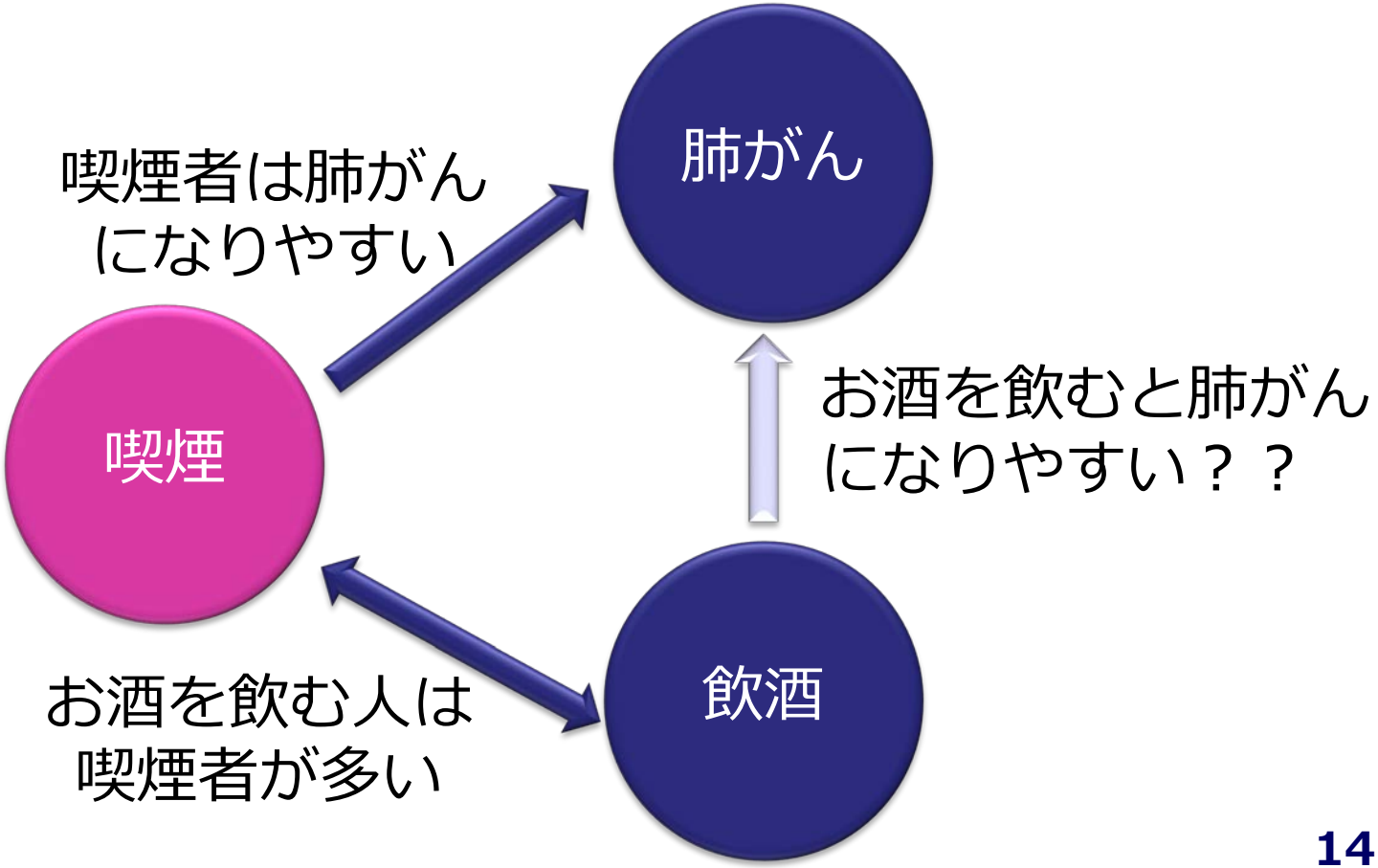
# 交絡についてのまとめ

- 治療法と予後に関連する第3の因子（年齢）によって見かけ上の関連が生じてしまう現象のこと
  - 交絡を引き起こす因子（=年齢）のことを**交絡因子**という



# その他の交絡の例

- 飲酒が肺がんを起こすわけではないのに、そう見えてしまう
  - この場合、喫煙が**交絡因子**



# 交絡がないことを保証するには

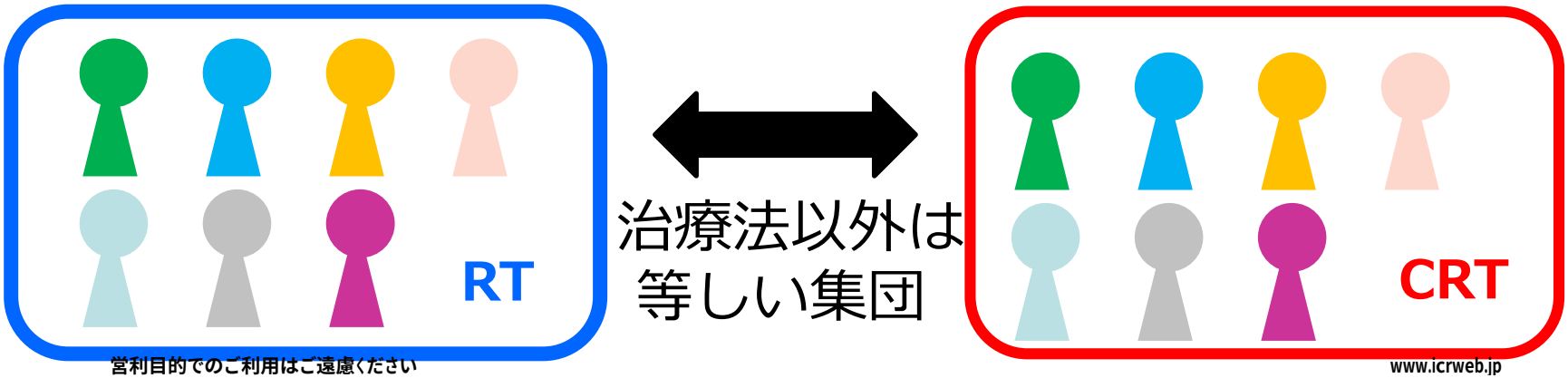
- 治療群間で予後に関係する背景因子を揃える
  - 年齢
  - Stage（がんの進行度）
  - Performance Status（全身状態）
  - その他（未知の因子も含めて）

因子がたくさんある・未知の因子があるために
全てを考慮できない

**ランダム**に決める

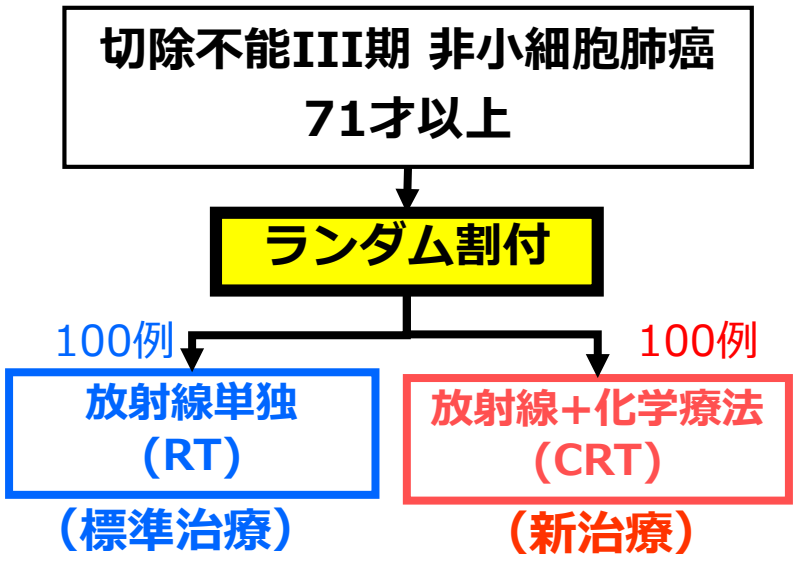
# ランダム化 randomization

- 医師あるいは患者の意思によらず、確率に基づいて各治療群に患者を割り付ける
- 予見による患者選択の偏りの防止
  - 状態の良い患者はCRTに割り付けられやすくなる、などを防ぐ
- **比較可能性（内的妥当性）が担保される**
  - 治療法以外は等しい集団 → 効果に差があれば治療法の違い

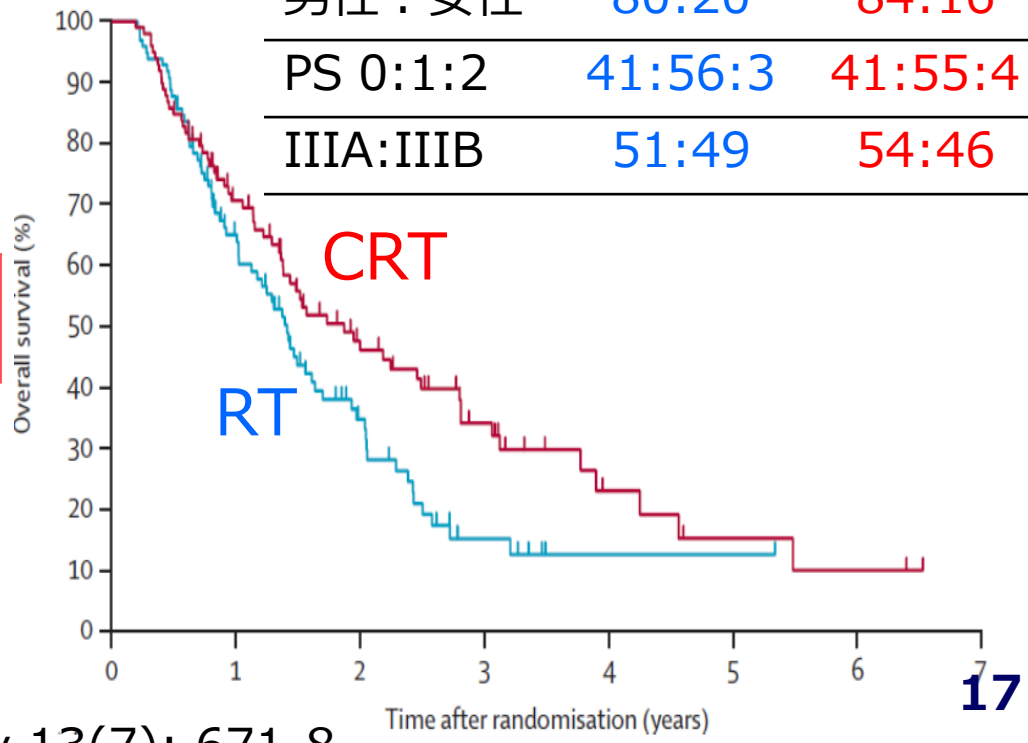


# JCOG0301の場合

- **RT**と**CRT**を比較するために**ランダムに割り付けた**
  - 治療法以外の背景因子は平均的に治療群間で同じ
  - 生存曲線の違いは治療法による違いであると期待できる



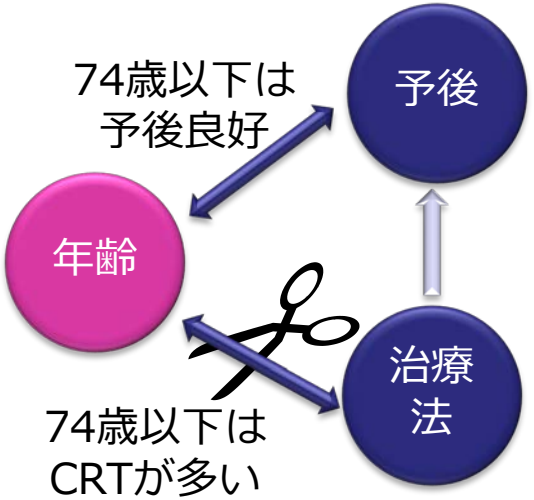

| | RT | CRT |
|---|---|---|
| 年齢中央値 | 77歳 | 77歳 |
| 男性：女性 | 80:20 | 84:16 |
| PS 0:1:2 | 41:56:3 | 41:55:4 |
| IIIA:IIIB | 51:49 | 54:46 |

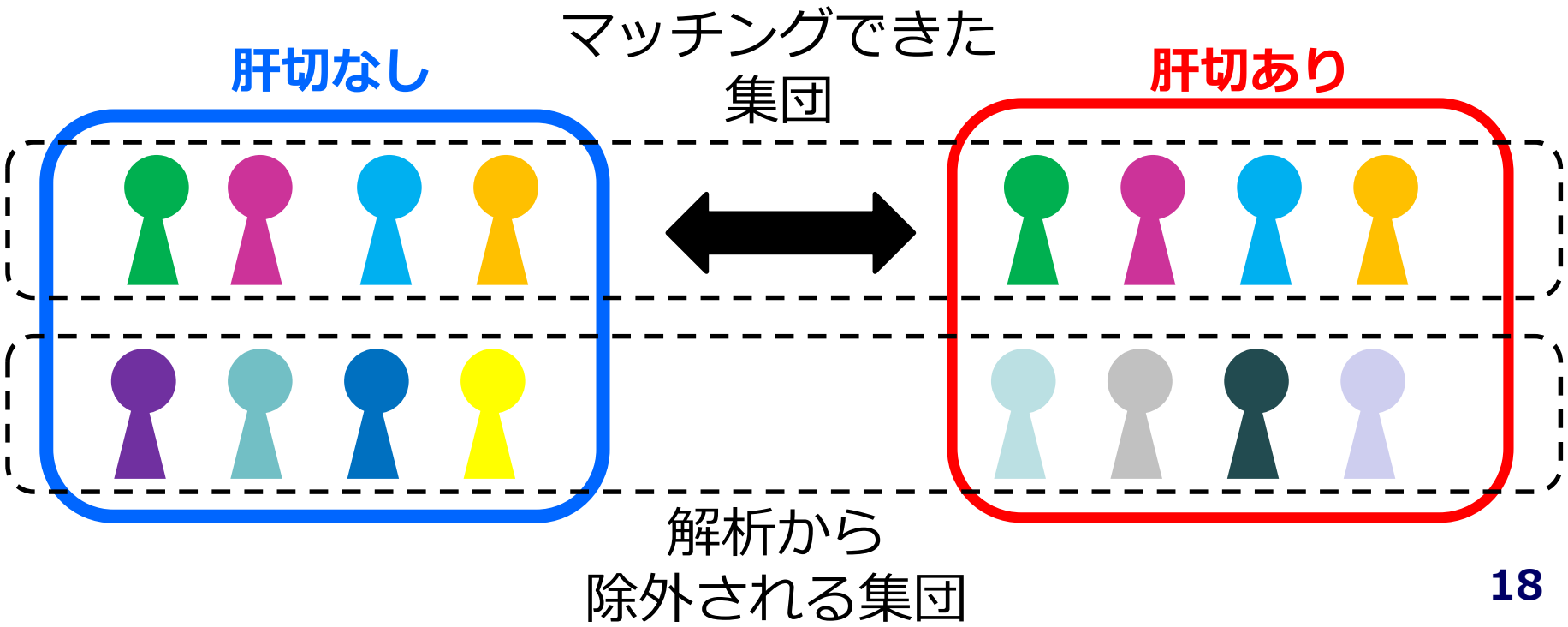



Atagi *et al.* (2012) Lancet Oncology 13(7): 671-8.

# 試験計画段階で対処する方法

- マッチング
  - アウトカムに強い影響を与えると考えられる背景因子が一致する患者どうしを合わせる



# 交絡への対処法

- 試験計画段階（ or データ集積）
  - ランダム化(randomization)
  - マッチング(matching)
- 解析段階
  - サブグループ解析(subgroup analysis)
    - サブグループ毎に治療効果を見る
  - 層別解析(stratified analysis)
    - サブグループ毎の結果を統合(重み付き平均)して、1つのp値、1つの治療効果を求める
  - モデルを用いた多変量解析(multivariate analysis)
    - ロジスティック回帰やCox回帰など
  - 傾向スコア
    - ある患者がある治療に割り付けられる確率を求め調整する

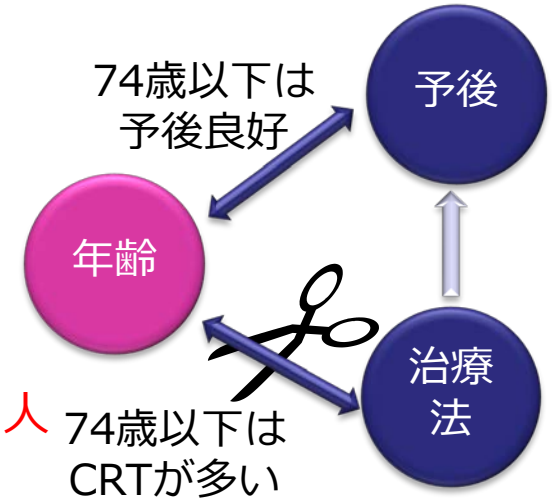
喫煙有無等、倫理的な側面からランダム化を伴う介入研究の実施が難しい場合がある

背景がマッチしない患者集団が解析対象から除外されてしまう。（選択バイアス？）

注：仮想例

# サブグループ解析

## 年齢別に治療と予後の関係を検討する




生存割合

**CRT** 190/250人

**RT** 40/60人

74歳以下

**CRT** 160/**200**人

**RT** 16/**20**人

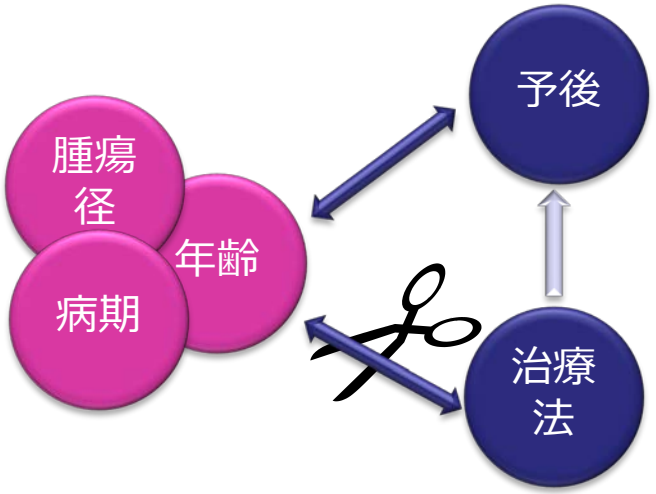
75歳以上

**CRT** 30/**50**人

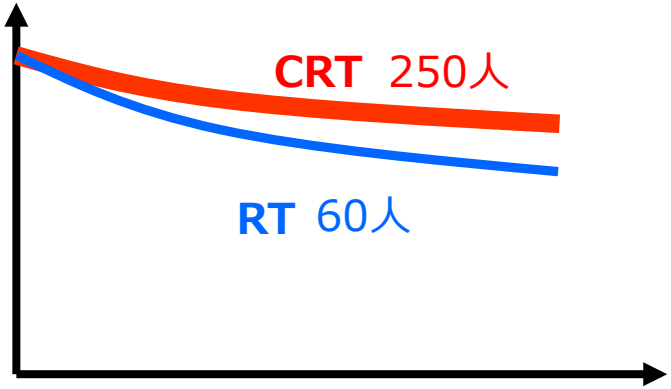
**RT** 24/**40**人

注：仮想例

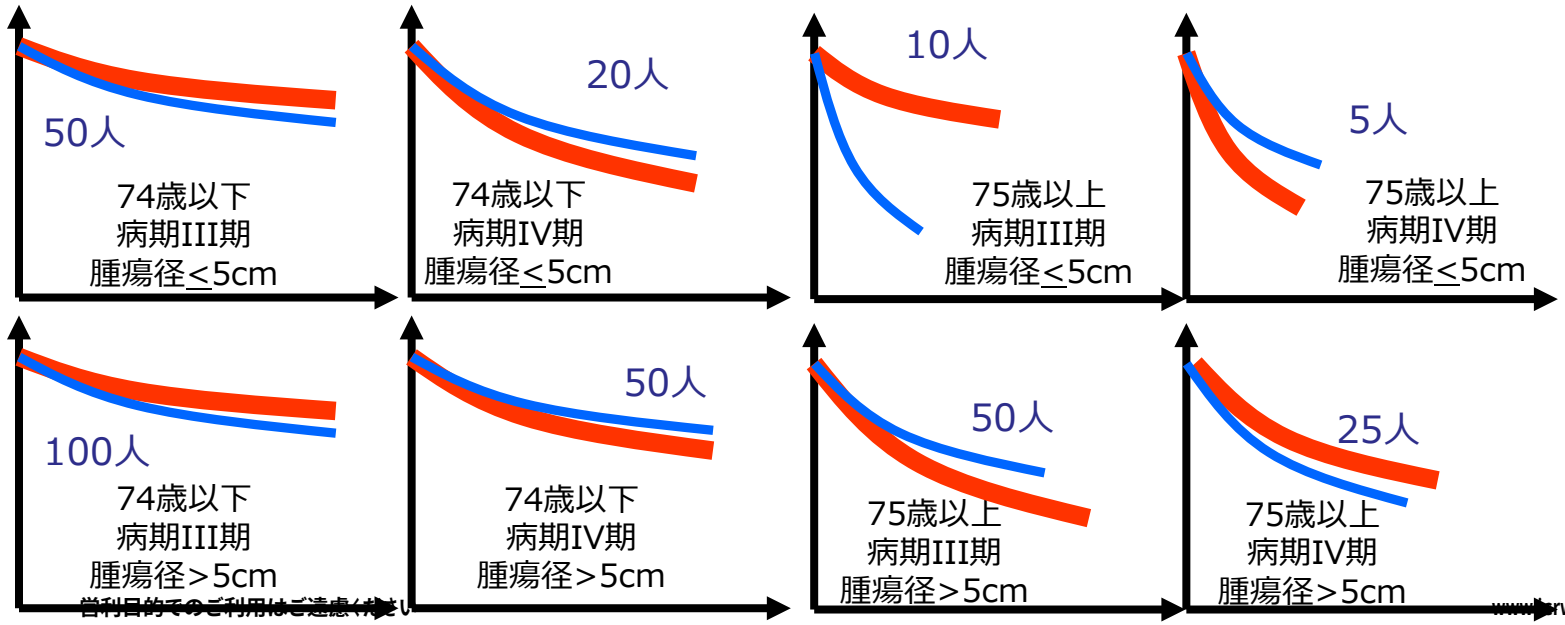
# サブグループ解析の欠点




## 年齢、病期、腫瘍径に関して治療と予後の関係を検討する




**交絡因子が複数の時、各サブグループのサンプルサイズが小さくなり過ぎる！**

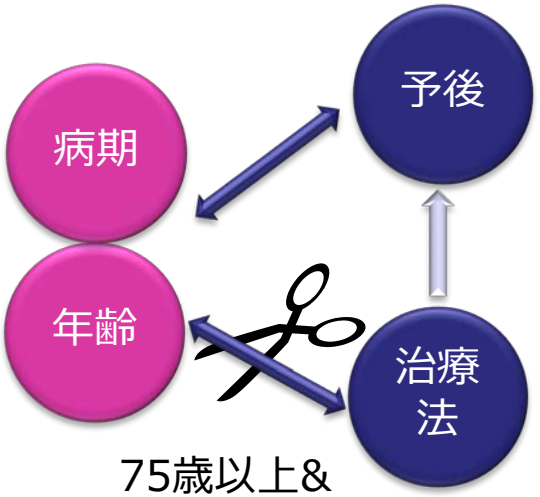


# サブグループ解析の利点・欠点

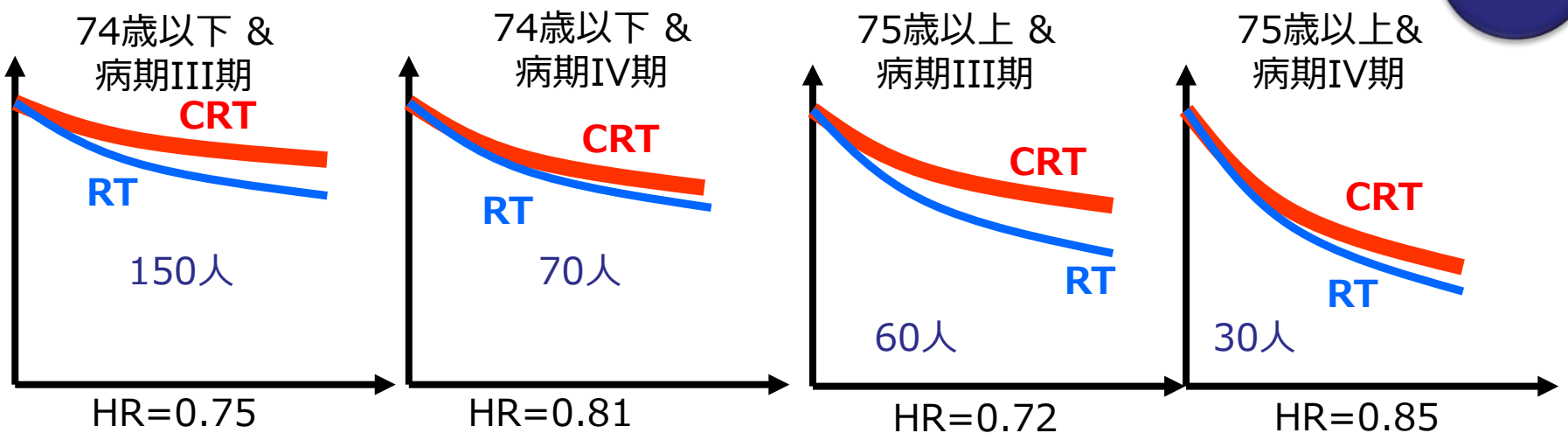
- 利点
  - 簡単でわかりやすい
    - 各サブグループで治療効果の検討をすればよいだけ
    - 統計的仮定が少ない
- 欠点
  - 集団全体での治療効果がわからない
  - サブグループが多すぎると、各サブグループのサンプルサイズが小さくなりすぎる
    - 交絡因子が5個あれば、少なくとも $2^5=32$個のサブグループ
    - 交絡因子が連続量の場合は、カテゴリー化してからでなければサブグループ解析できない
  - 交絡因子自体(年齢が74歳以下に対する75歳以上)の効果の大きさがわからない

注：仮想例

# 層別解析(stratified analysis)




## 年齢、病期のサブグループ別に求めた治療効果を統合




## 個々のサブグループの治療効果(HR)を重み付き平均して1つに統合

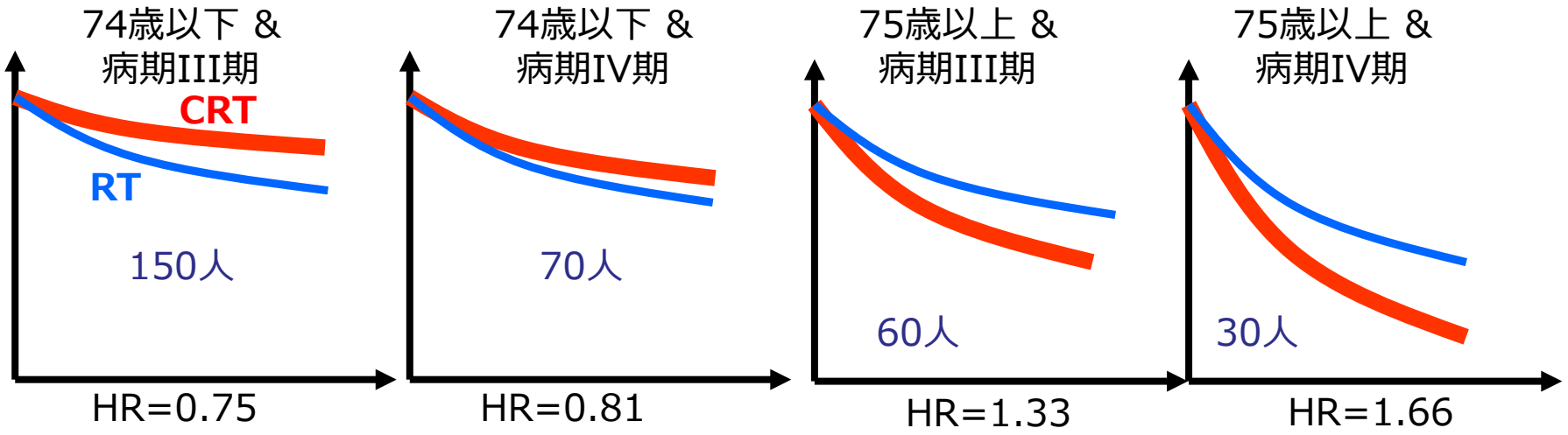
$$\text{全体の効果}(HR) = 0.75\times\frac{150}{310} + 0.81\times\frac{70}{310} + 0.72\times\frac{60}{310} + 0.85\times\frac{30}{310} = 0.77$$

重みには、サンプルサイズや推定値のバラツキなどを用いる

注：仮想例

# 層別解析の前提

各サブグループで治療効果が同程度だとみなせること




$$\text{全体の効果}(HR) = 0.75 \times \frac{150}{310} + 0.81 \times \frac{70}{310} + 1.33 \times \frac{60}{310} + 1.66 \times \frac{30}{310} = 0.96(??)$$

# 層別解析の利点と欠点

- 利点
  - 集団全体の治療効果を求めることができる
  - (モデルを用いた解析と比較して) 仮定が少ない
- 欠点
  - サブグループが多すぎると、各サブグループのサンプルサイズが小さくなりすぎる
    - 交絡因子が5個あれば、少なくとも $2^5$=32個のサブグループ
    - 交絡因子が連続量の場合は、カテゴリー化してからでなければ解析できない
  - 交絡因子自体(年齢が74歳以下に対する75歳以上)の効果の大きさがわからない

# 再掲：交絡への対処法

今回紹介した方法

- 試験計画段階（ or データ集積）
  - ランダム化(randomization)
  - マッチング(matching)

簡単だが、同時に扱える因子の数に制限がある.
連続量の因子が扱えない

- 解析段階
  - サブグループ解析(subgroup analysis)
    - サブグループ毎に治療効果を見る
  - 層別解析(stratified analysis)
    - サブグループ毎の結果を統合(重み付き平均)して、1つのp値、1つの治療効果を求める
  - モデルを用いた多変量解析(multivariate analysis)
    - ロジスティック回帰やCox回帰など
  - 傾向スコア
    - ある患者がある治療に割り付けられる確率を求め調整する

# 仮説検定

# CRT群は勝ったの？

肺がん内科グループ
JCOG0301

**切除不能III期 非小細胞肺癌**
**71才以上**

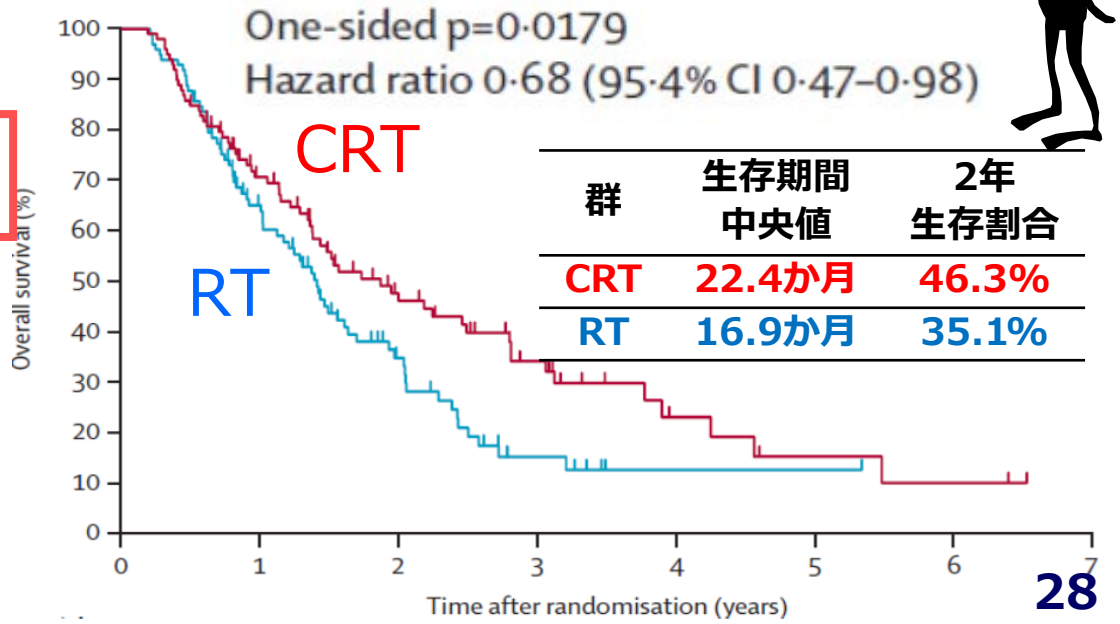
↓

**ランダム割付**

100例 → **放射線単独（RT）** **（標準治療）**

100例 → **放射線+化学療法（CRT）** **（新治療）**

ランダム化しているから比較可能性があることはわかった。確かに、CRT群の生存曲線がRT群よりも上にあるけど、ランダム化して生存曲線が上にあればCRT群が勝ったと言って良いの？



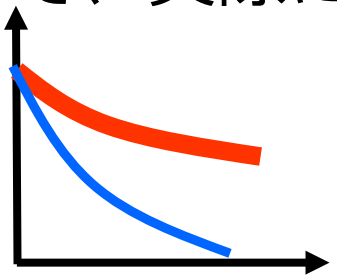

| 群 | 生存期間中央値 | 2年生存割合 |
|---|---|---|
| CRT | 22.4か月 | 46.3% |
| RT | 16.9か月 | 35.1% |



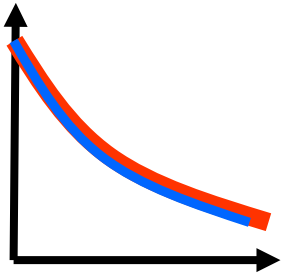
Atagi *et al.* (2012) Lancet Oncology 13(7): 671-8.

# 生存曲線が開いている時の解釈

- **2つの可能性がある。どちらが正しい？**
  - 本当に「**RT**と**CRT**に差がある」ので、実際に差が出た
    - 正しい結論を得ている



  - 本当は「**RT**と**CRT**に差がない」のに、偶然差があるように見えた
    - 誤った結論をしてしまっている



**どちらが正しいか、得られた結果から確かめたい！**

# 確かめる方法：仮説検定

- 証明したいことは、「RTとCRTに差がある」 ですが、

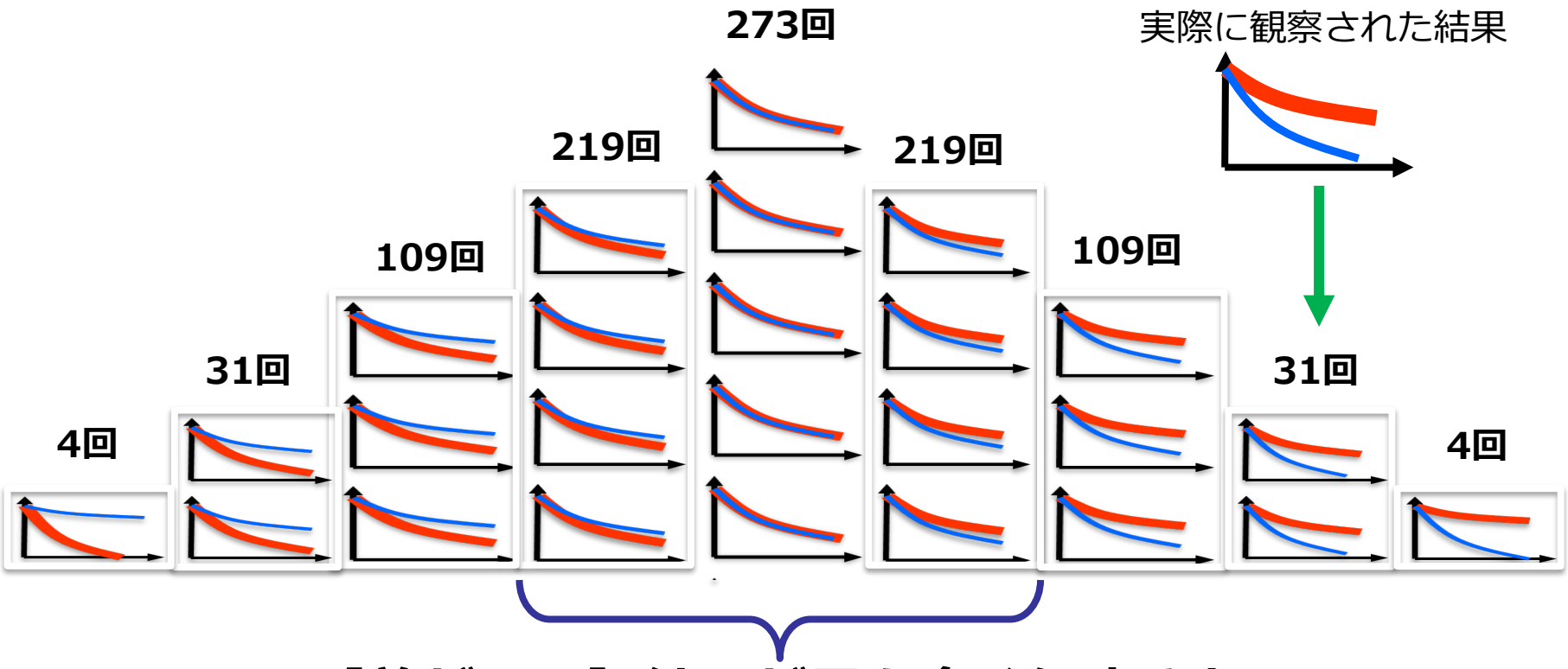
1. 「RTとCRTに差がない」という仮説を置く
   - この仮説を**帰無仮説**という
2. 「RTとCRTに差がない」という仮定の下で、何回も試験をした場合に得られる結果の分布を調べる
3. 実際に観察されたRTとCRTの差以上に大きな差になる確率を調べる
4. この確率が小さければ、そもそも「RTとCRTに差がない」という仮説（帰無仮説）が間違っていた、と判断する
5. 「RTとCRTに差がある」が正しいと判断する

# RTとCRTの生存曲線に【差がない】下での結果の分布

もし、RTとCRTの生存曲線に【差がない】が真実なら…

日本全国の切除不能III期 非小細胞肺癌

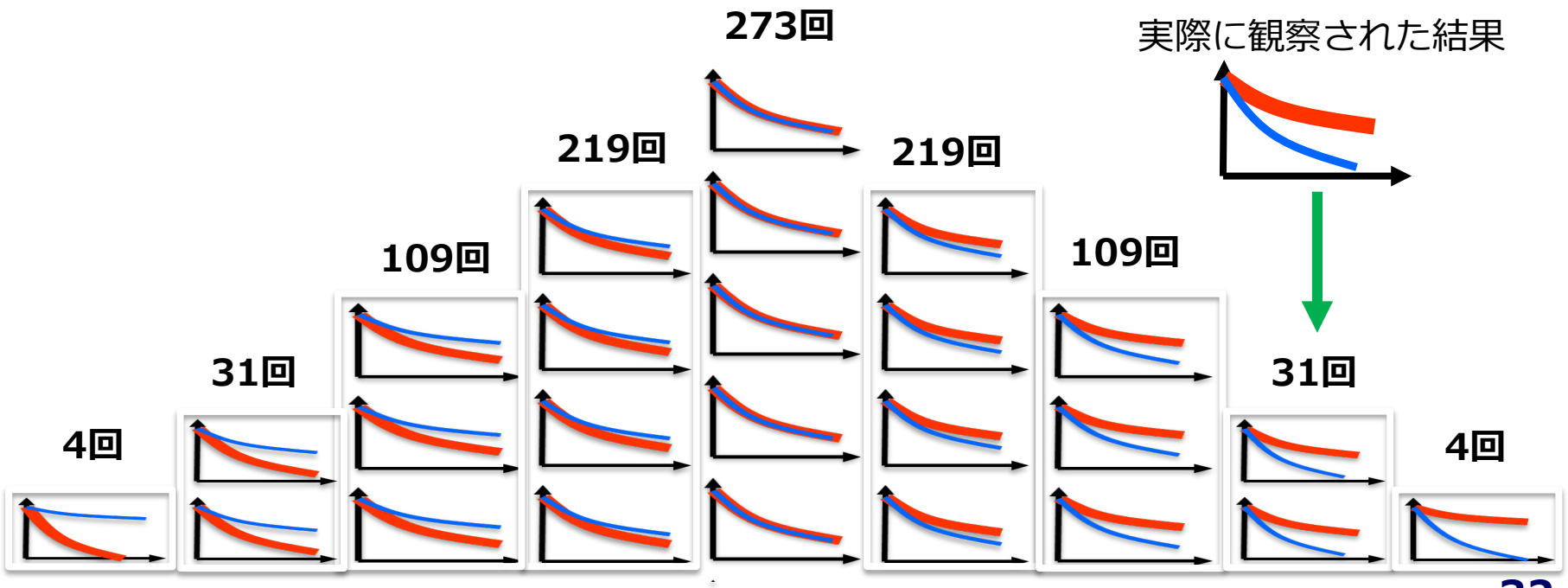
71才以上の患者から200人選んで1000回試験すると、、、




**【差がない】結果が最も多く観察される**

# P値の計算

- 実際に観察された結果以上に大きな差になる確率(**P**robability)は、35/1000 = **3.5%**
  - この確率のことを**p値**という
- 実際に観察された結果は【差がない】が真実だとしたら、1000回中35回くらいしか起こらないような**稀な結果**（?）



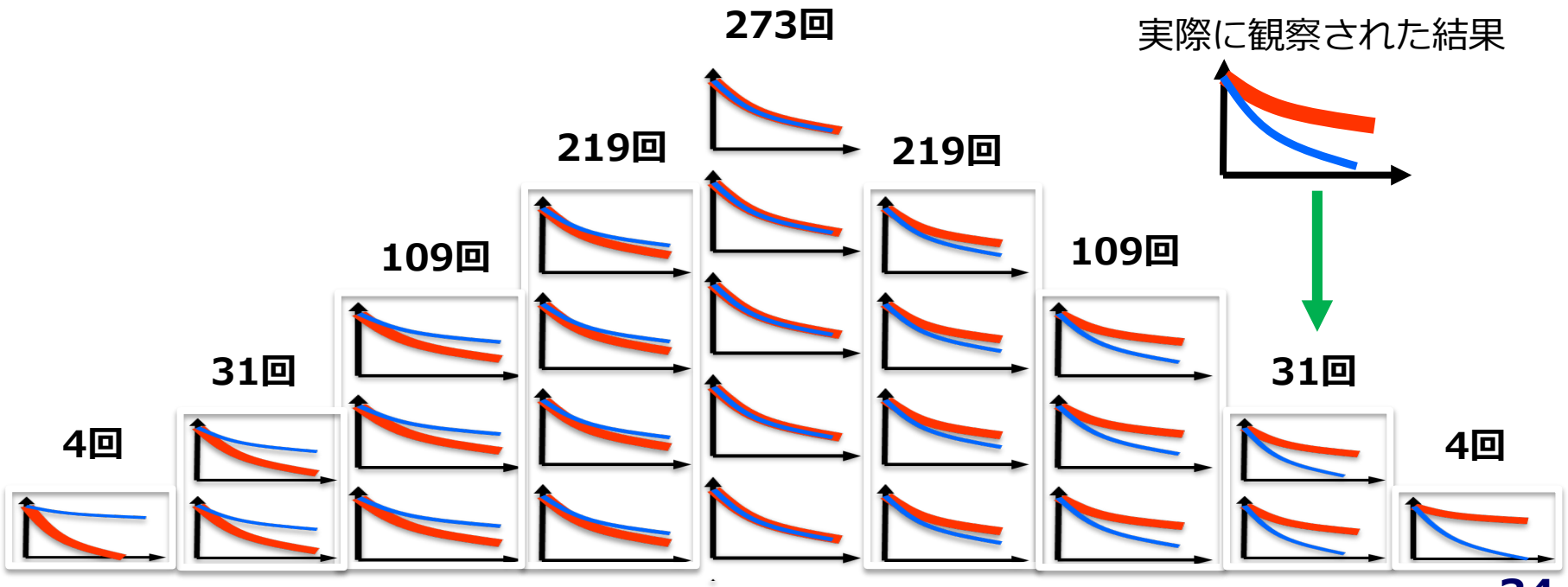
# 3.5%は稀な結果？

- 3.5%は**稀な結果と考える**場合
  - そもそも【差がない】という仮説が間違っていたと判断し、RTとCRTは**差があると結論する** ＝ **【有意差あり】**
- 3.5%は**稀な結果とは考えない**場合
  - 【差がない】という仮説は間違っているとは言えないので、RTとCRTに**差があるとは言えないと結論する** ＝**【有意差なし】**

- 結果を見てから稀かどうかを判断すると後付けになってしまうので、事前に稀かどうかの規準を決めておく
  - この規準のことを**有意水準(α level)**という
  - p値が有意水準を下回ったら、【有意差あり】と結論する

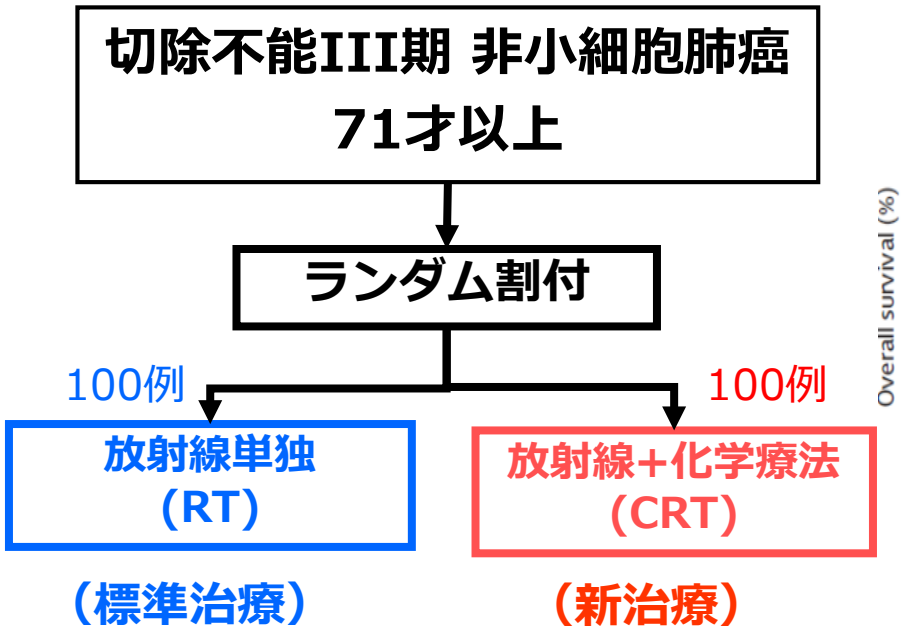
# 検定結果

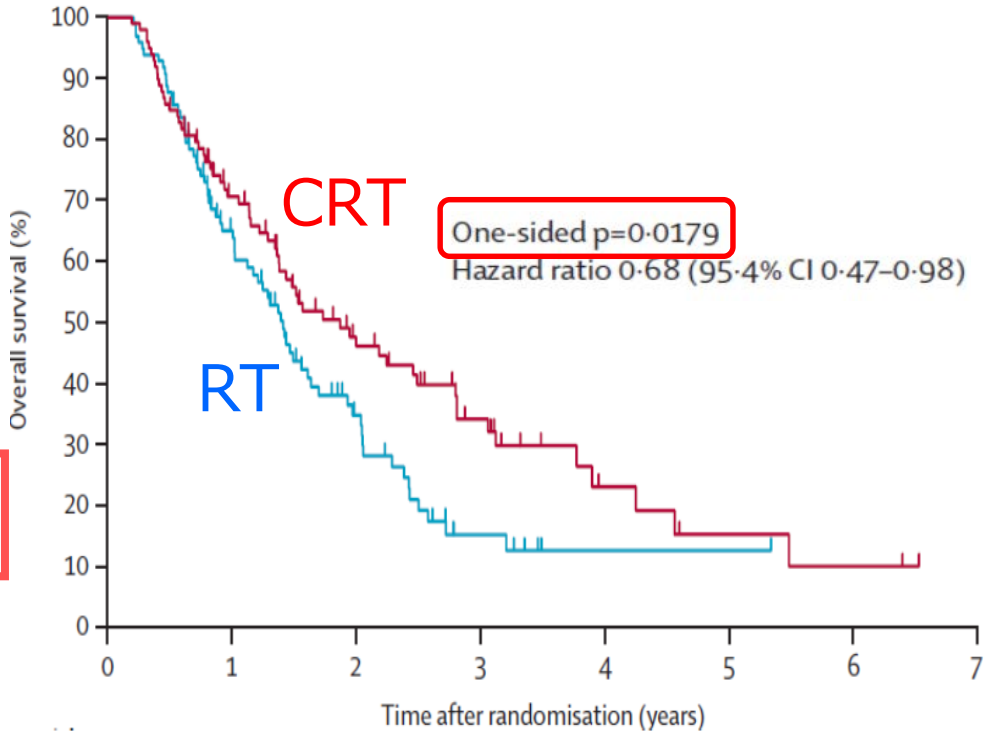
- P値=3.5%だった
  - 実際に観察された結果は【差がない】が真実だとしたら、1000回中35回くらいしか起こらないような結果
- 有意水準を**5%**に設定していたとしたら、有意差あり
- 有意水準を**2.5%**に設定していたとしたら、有意差なし



# JCOG0301の場合

- **<u>p=0.0179</u>**：両群に差がないとしたら100回中1～2回くらいしか起こらない稀な事象
  - 事前に決めた規準(有意水準) α≦5%も満たす
  - CRTはRTと比較して優れていると判断






Atagi *et al.* (2012) Lancet Oncology 13(7): 671-8.

# αエラー、βエラー、検出力

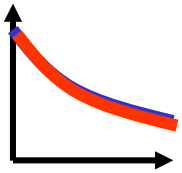
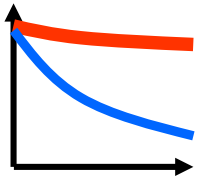
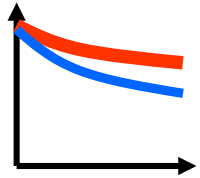
仮説検定の結果は絶対正しい？？

# 必ずしも検定の結果は正しいとは限らない

- 実際に得られた結果はP値=3.5%
  - 有意水準5%とすると、稀にしか起こらない事象なので、【差がない】という仮説は誤っていると判断した
  - 逆に言えば、【差がない】が真実の場合に稀には起こる
- 真実が【差がない】時に、誤って【差がある】と判断してしまうのは誤った判断をしていることになる
  - この誤りのことを**αエラー**という
  - 【差がない】時に【差がある】と判断する確率は有意水準以下なので、αエラーを起こす確率は有意水準以下

# 【差がある】のに有意差なしとしてしまう

- この誤りのことを、**”β エラー”**と呼ぶ
  - 本当は効果がある治療を真実に反して捨ててしまう誤り
- **検出力**（確率は**1-β**）
  - 「差がある」ものを正しく「差がある」と判断する確率

| | | 真実 | |
|---|---|---|---|
| | | 帰無仮説（差がない）  | 対立仮説（差がある）  |
| 検定結果  | 有意差なし | 正しい | 誤り<br>（**βエラー**） |
| | 有意差あり | 誤り<br>（**αエラー**） | 正しい<br>（**検出力.1-β**） |

# 治療効果の推定
# p値ではわからないこと

# CRTはどのくらい良い治療？

肺がん内科グループ
JCOG0301

**切除不能III期 非小細胞肺癌**
**71才以上**

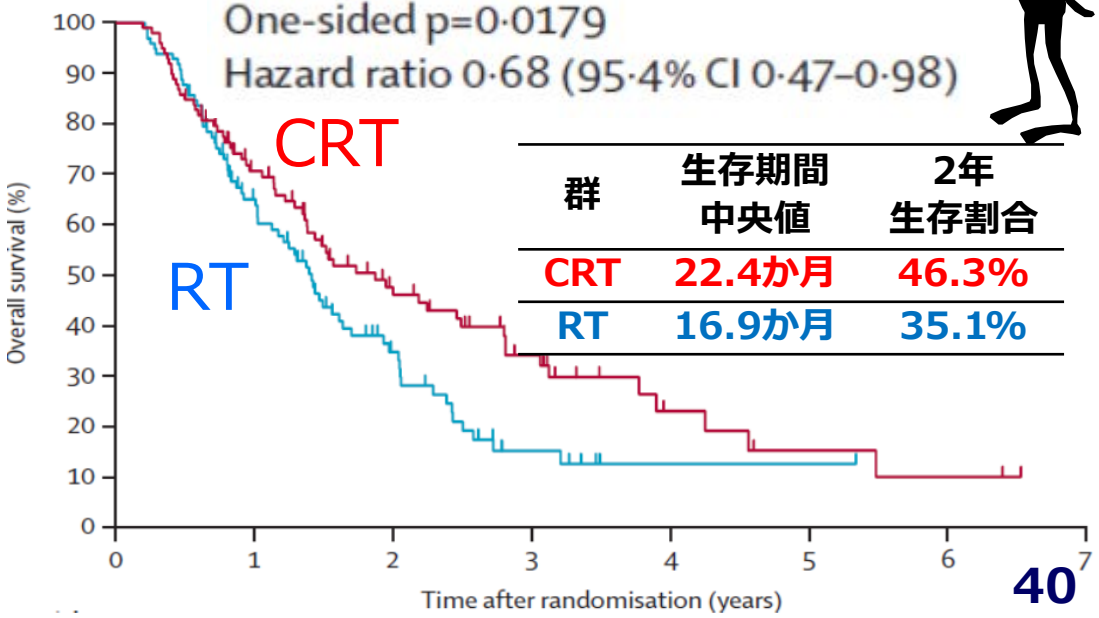
↓

**ランダム割付**

100例 → **放射線単独 (RT)** **（標準治療）**

100例 → **放射線+化学療法 (CRT)** **（新治療）**

CRT群がRT群よりも良いことはわかった。
でも、どのくらい良い治療法なの？
P値が小さければ良い治療？




| 群 | 生存期間 中央値 | 2年 生存割合 |
|---|---|---|
| CRT | 22.4か月 | 46.3% |
| RT | 16.9か月 | 35.1% |



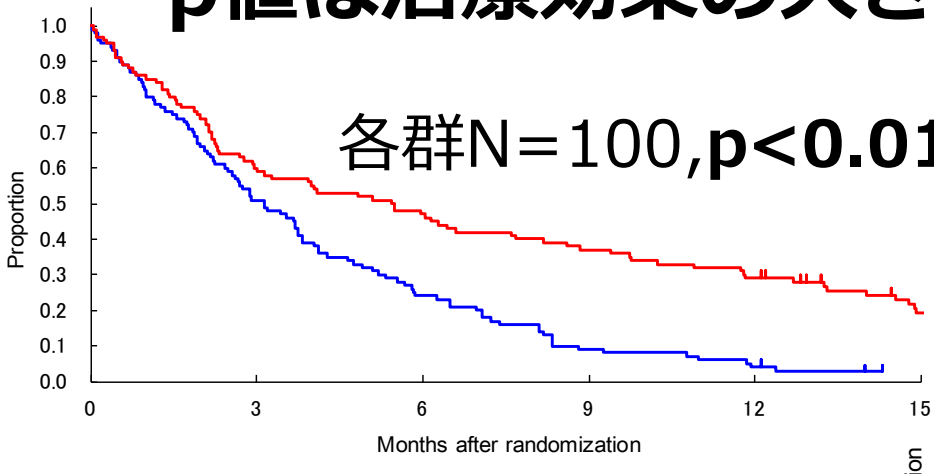
Atagi *et al.* (2012) Lancet Oncology 13(7): 671-8.

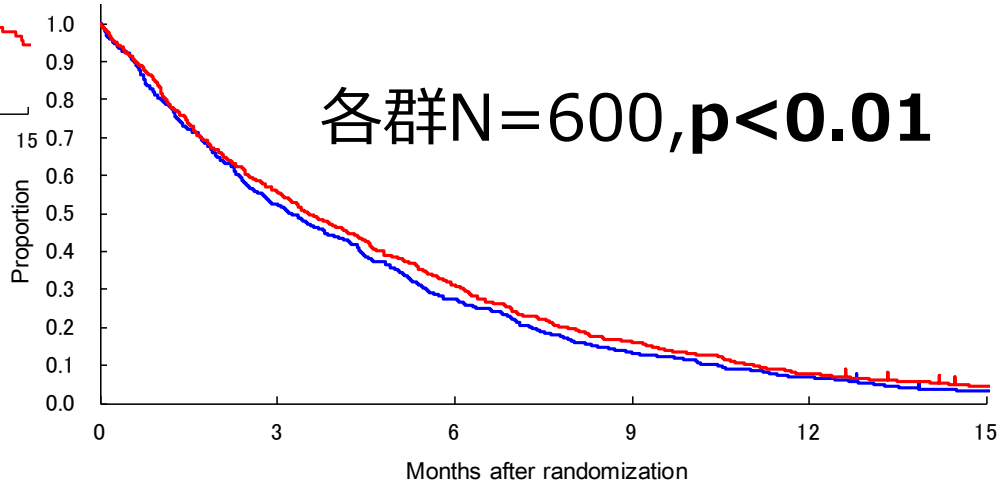
注：仮想例

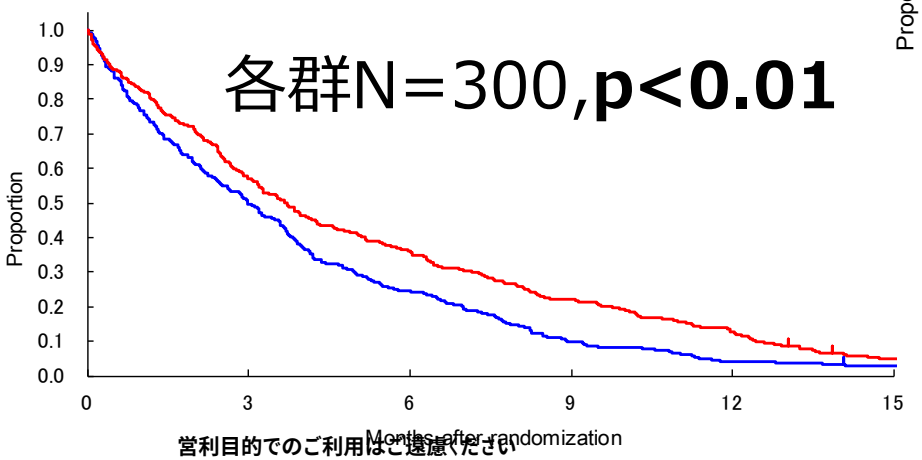
# 統計的有意差 ≠ 臨床的有意差

## 同じp<0.01でも臨床的意味は異なる

## p値は治療効果の大きさを表す指標ではない



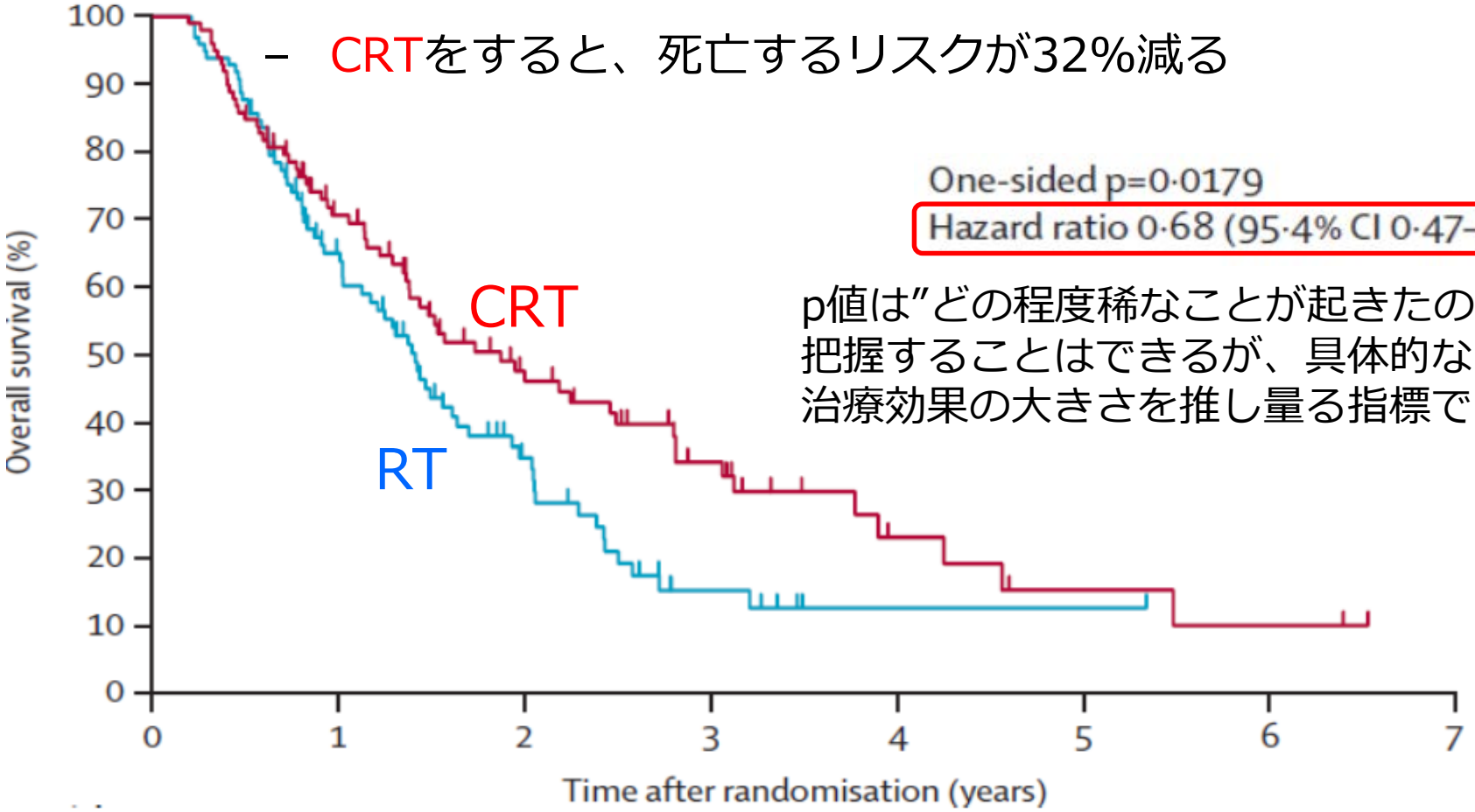



# 治療効果の大きさを表す指標

- 曲線のある1時点に着目した指標
  - 時点生存割合
    - 2年生存割合 CRT:46.3% vs RT:35.1%
  - 生存期間中央値（MST）
    - CRT: 22.4か月 vs RT:16.9か月
- 曲線全体を一つの効果にまとめた指標
  - **ハザード比**(**HR**：Hazard Ratio)
    - 群間の**ハザード（瞬間死亡率）**の比をとったもの

# JCOG0301における解釈

- RT群に対するCRT群の**ハザード比(HR)が0.68**
  - CRTをすると、死亡するリスクが0.68倍になる
  - CRTをすると、死亡するリスクが32%減る




p値は”どの程度稀なことが起きたのか”を把握することはできるが、具体的な治療効果の大きさを推し量る指標ではない

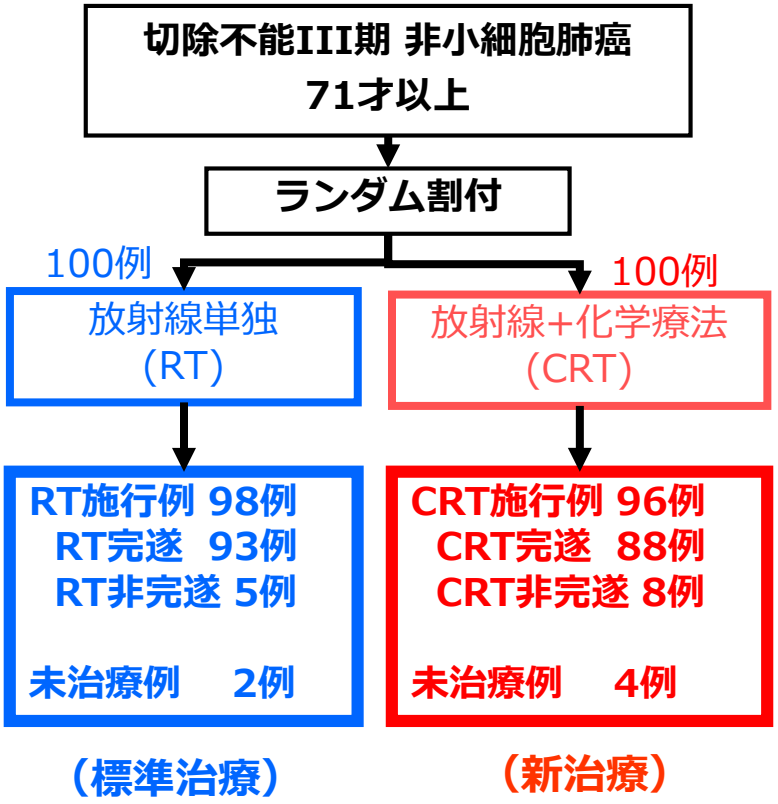


Atagi *et al.* (2012) Lancet Oncology 13(7): 671-8.

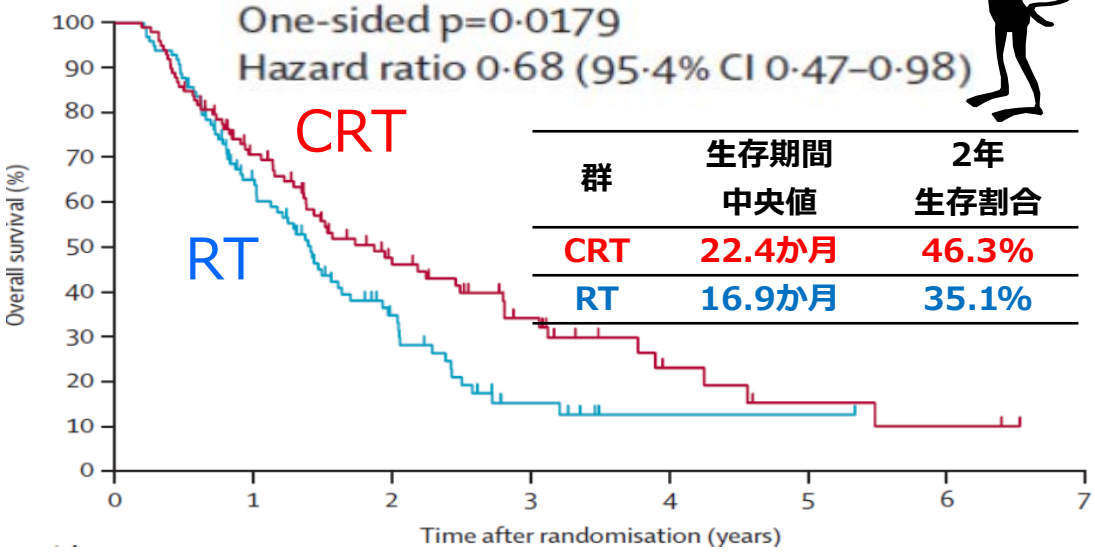
# Intent(ion)-To-Treat

# ちゃんと治療していない患者の扱いは？

肺がん内科グループ
JCOG0301




RT群にもCRT群にも、ちゃんと治療を出来てない人がいるぞ！？
この患者は生存曲線の群間比較の解析に含まれているのかな？

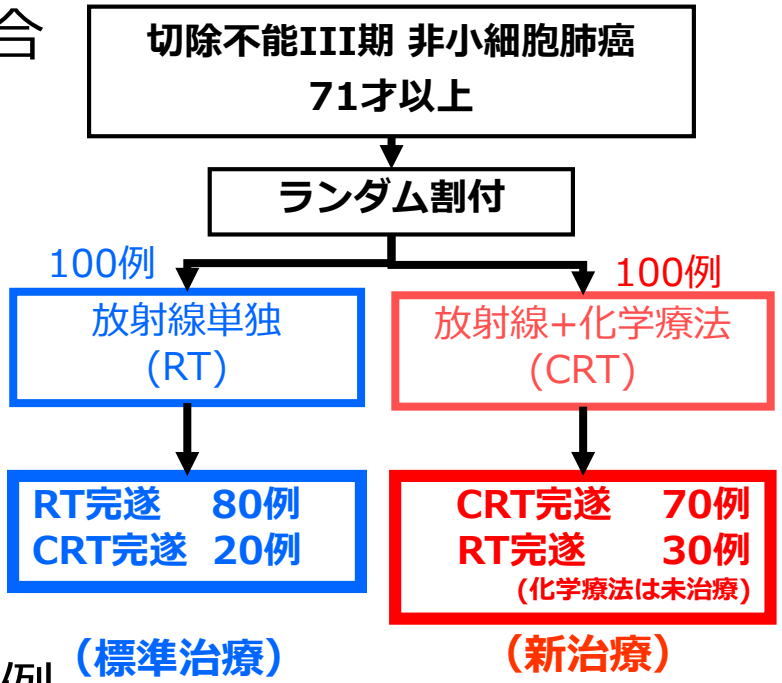



| 群 | 生存期間 中央値 | 2年 生存割合 |
|---|---|---|
| CRT | 22.4か月 | 46.3% |
| RT | 16.9か月 | 35.1% |

Atagi *et al.* (2012) Lancet Oncology 13(7): 671-8.

# あなたならどの解析法を選びますか？

- 簡単のために右のような状況の場合どの解析法を選ぶのが良い？




**① 治療完遂例どうしで比較**

RT 80例 vs. CRT 70例

**② 実際に行った治療同士で比較**

RT (80+30)例 vs. CRT (70+20)例

**③ ランダム化で割り付けられたどうしで比較**

RT 100例 vs. CRT 100例

# どんな結果が予測されるか考えると

差がないのに誤って差があると言ってしまう確率≦p

① **治療完遂例どうしで比較**　RT 80例 vs. CRT 70例

- 治療完遂できない人は予後の悪い人なので、残った人は元気な人

② **実際に行った治療同士で比較** RT (80+30)例 vs. CRT (70+20)例

- CRTに割り付けられたけどRTしか出来なかった人は予後の悪い人
  RTに割り付けられたけどCRTをした人は元気な人

⇒　①、②の比較では群間で背景因子が揃わなくなり、
ランダム化した意味がなくなってしまう

差がない時は正しく差がないと言える

③ **ランダム化で割り付けられたどうしで比較** RT 100例 vs. CRT 100例

- 本当にCRTに効果がある場合、CRTに割り付けられたけどRTをした人がいると、治療効果が薄まってしまう

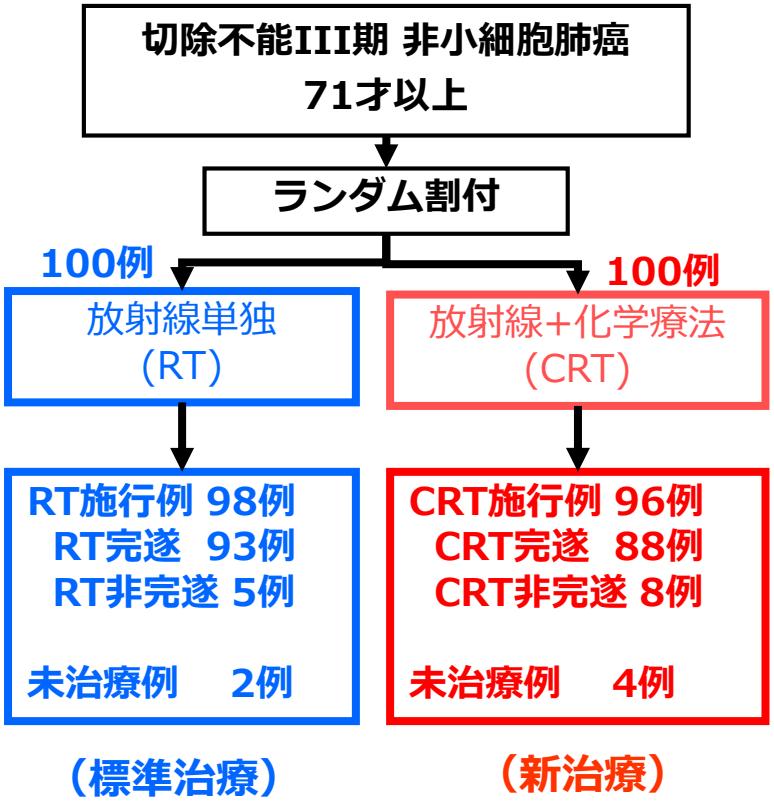
# Intention-to treat解析(ITT解析)

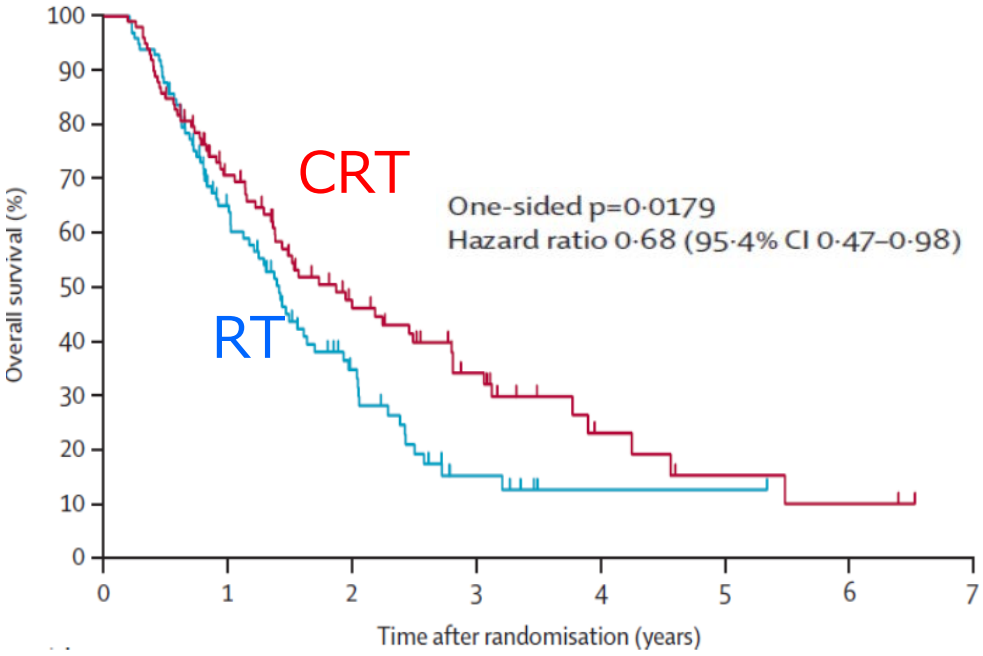
- ランダム化によって**割り付けられた通り**の治療群で行う解析（③の対象で解析する方法）のこと
  - **ITT解析**をすれば、αエラーを起こす確率は大きくならない
  - 有意になりにくいという意味で「保守的(conservative)」な解析法

- ITT解析でも有意差があったのなら、自信を持って差があると判断できる
  - 「少なくともXXXくらいの治療効果があります！」と言える
  - **ITT解析が臨床試験の主たる解析の標準的方法**

# JCOG0301の場合

- 主たる解析は治療非完遂例などを含めた**ITT解析**
  - CRTはRTに上回っていると判断できる




Atagi *et al.* (2012) Lancet Oncology 13(7): 671-8.

# まとめ

- **生存曲線**は生存割合を時間に対してプロット。結果を視覚的に判断できる
- **ランダム化**や**マッチング**によって交絡を除去し、比較可能性を担保
- **サブグループ解析**や**層別解析**によって、解析段階で交絡を除去可能
- 結果は仮説**検定**によって求めた**p値**が**有意水準（α）**を下回ったら差があると判断
- 治療効果の大きさはp値ではなく、**ハザード比**や生存割合で判断
- αエラーを制御するため**ITT解析**がランダム化試験の主たる解析法

# 参考図書


浜田知久馬
真興交易医書出版部; 新版
ISBN-13: 978-4880038612


佐藤俊哉
岩波書店
ISBN-13: 978-4000074544


大橋靖雄
医歯薬出版
ASIN: B00I399SY4


佐藤俊哉
岩波書店
ISBN-13: 978-4000295949