

動機(1/2)

- 意思決定者: 今回の第III相試験(Ph3)が成功する確率は?
- 統計担当: いつも90%です
 - 条件: 第II相試験(Ph2)の結果に基づいて, 検出力90%で試験デザインを組んでいるときに, 真の薬効が想定通りだとした場合

- Ph2の試験規模(サンプルサイズ)が大きいか小さいか?
- Ph2の評価項目(エンドポイント)がPh3と同じか違うか?
- そもそも第II相を実施しているかどうか?

- に依存しない?

動機(2/2)

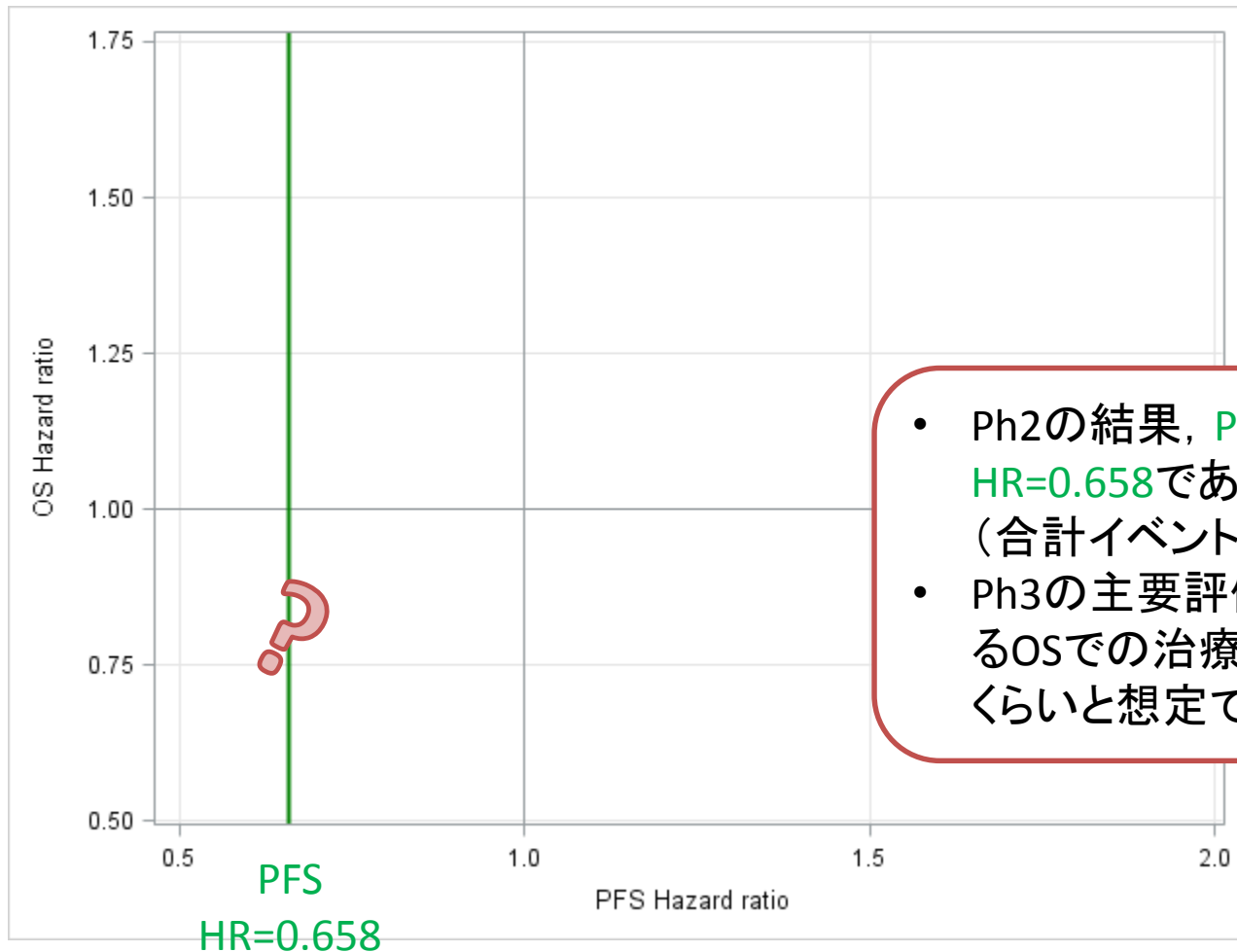
- 臨床開発におけるPh2の目的
 - 様々な用量の安全性・忍容性の評価する
 - 有望そうな用量幅を選択する
 - 化合物による早期のサインを見つける
 - Ph3でどのくらいの効果が見込めるかを見極める
 - 疾患領域によっては,
accelerated approval(FDA)/conditional approval(EMA)に利用
- Ph3の治療効果の見極め, どうやっていますか？

事例1: 膵臓がん(Ph2: PFS, Ph3: OS)

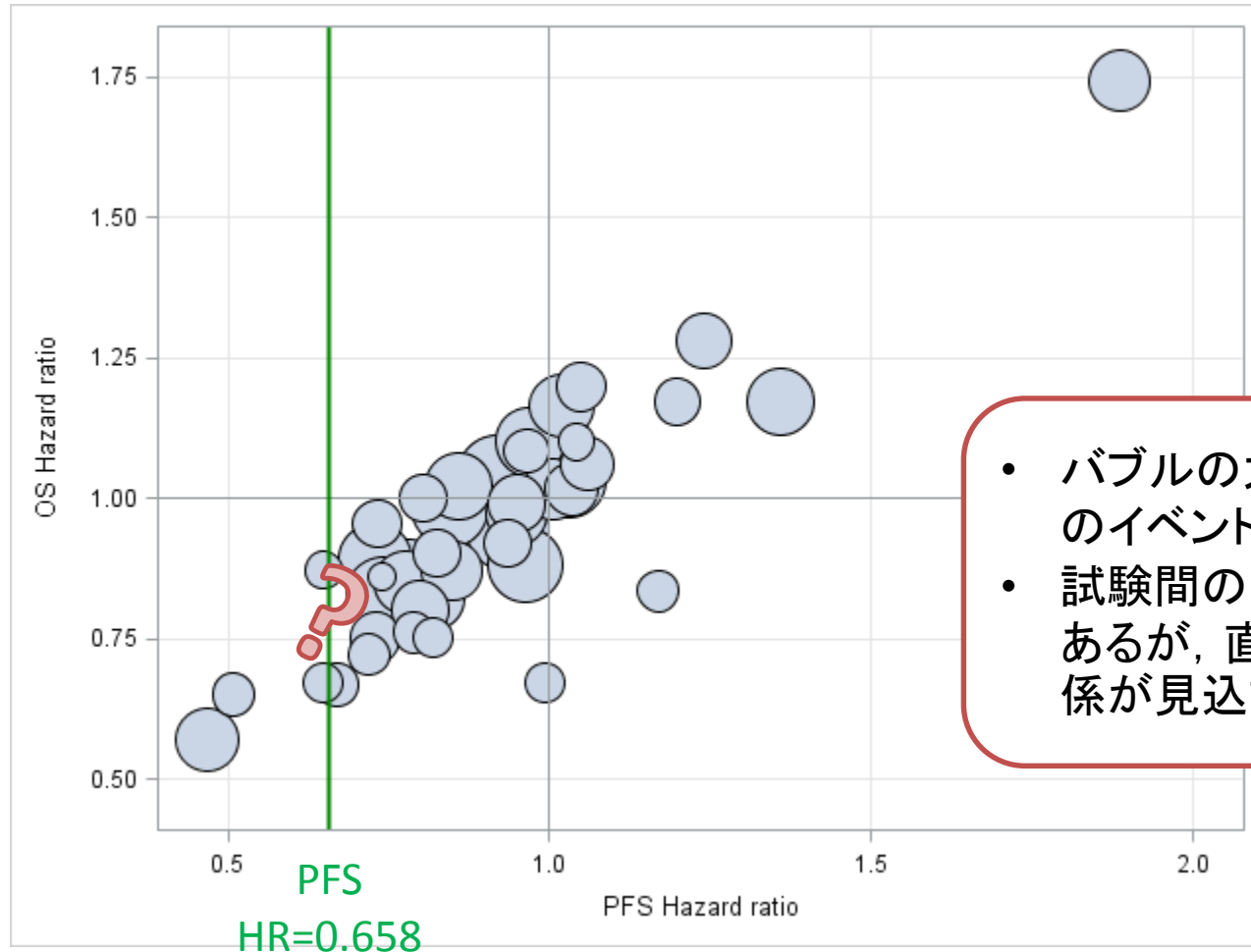
- Ph2とPh3で評価項目が異なる場合に, Ph2の結果からPh3の治療効果を見極める方法として, ベイズ流アプローチを用いた事例
 - Sabin et al. (2014), Sabin et al. (2015)を参考
 - Amgenで用いられている(論文当時の情報)
- ディスカッションポイント
 - Ph2の結果からPh3の治療効果を見極める方法(今回はベイズ流メタ回帰)は有用か?
 - 治療効果を固定した検出力に対して, 治療効果のばらつきを踏まえて検出力の期待値をとったPoS (Probability of Success)*という考え方は有用か?

* なお, successは有意/非有意以外にも設定可能であるため, 検出力の期待値をとったものだけがPoSというわけではない

Ph2の結果が得られたとします (RCT, gemcitabine単剤との比較, 1st line)

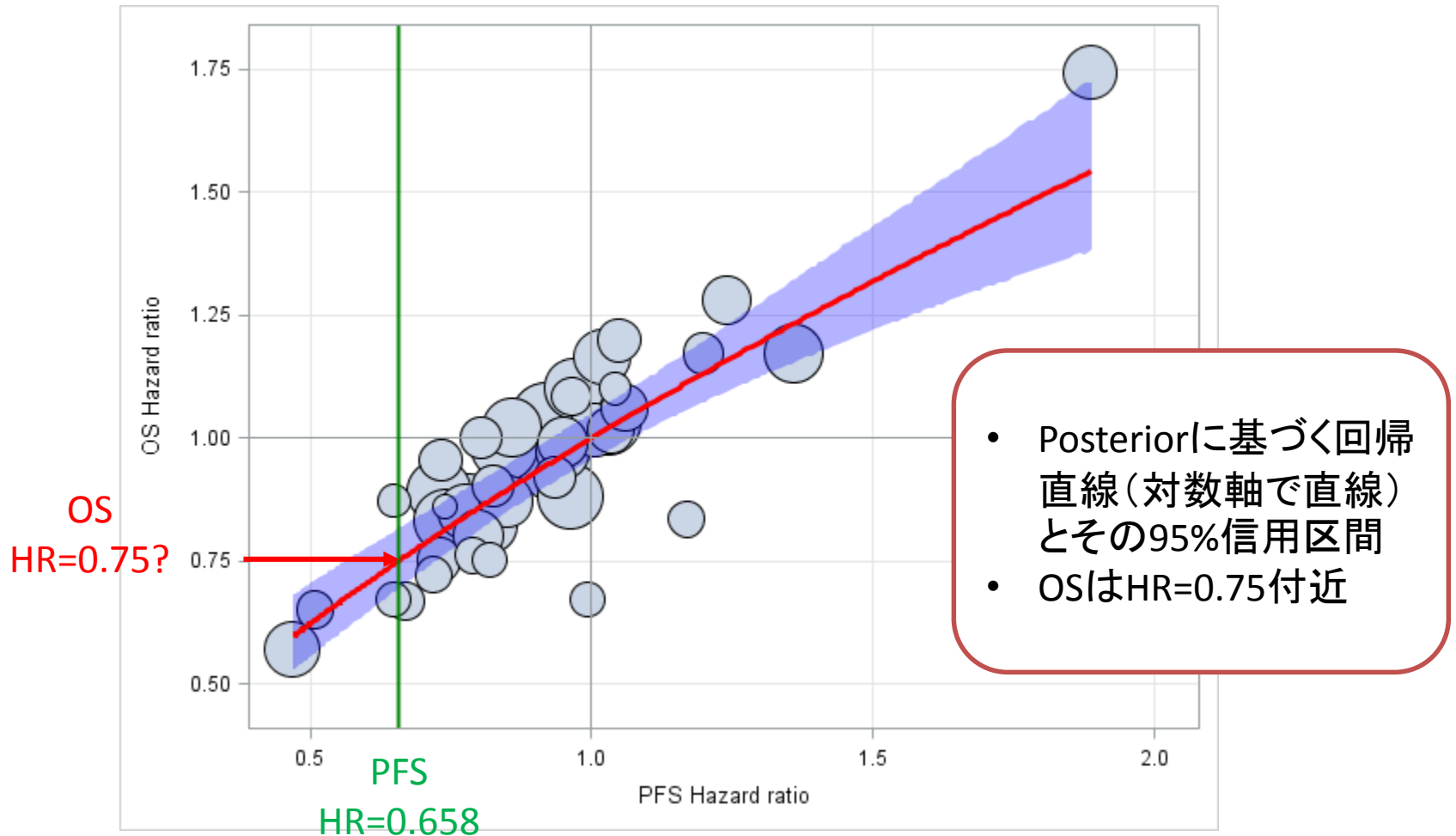


システマティックレビューで得た文献データをプロット (2000-2012年の膵臓がんに対するRCT, gemcitabine単剤との比較, 1st line)



- バブルの大きさはOSのイベント数に比例
- 試験間のばらつきはあるが、直線的な関係が見込まれる

過去のエビデンスに基づくOSに対する治療効果は？



対数ハザード比に対するベイズ流メタ回帰

緑: data
赤: parameter

モデル式

- $\hat{\theta}_i \sim N(\mu_i, \epsilon_i^2)$
- $\mu_i \sim N(\phi_i, \tau^2)$
- $\phi_i = \beta \times Z_i$ (論文は $\mu_i = \beta \times Z_i$)

- $\hat{\theta}_i$: i番目の試験で観測されたOSのln HR
- ϵ_i : 上記の標準誤差(真とする)
- μ_i : i番目の試験の真のOSのln HR
- ϕ_i : PFSから得られる真のOSのln HR
- τ^2 : 試験間のばらつき
- β : 傾き(OSとPFSの関係性)
- Z_i : i番目の試験で観測されたPFSのln HR (PFSは定数と考えている)

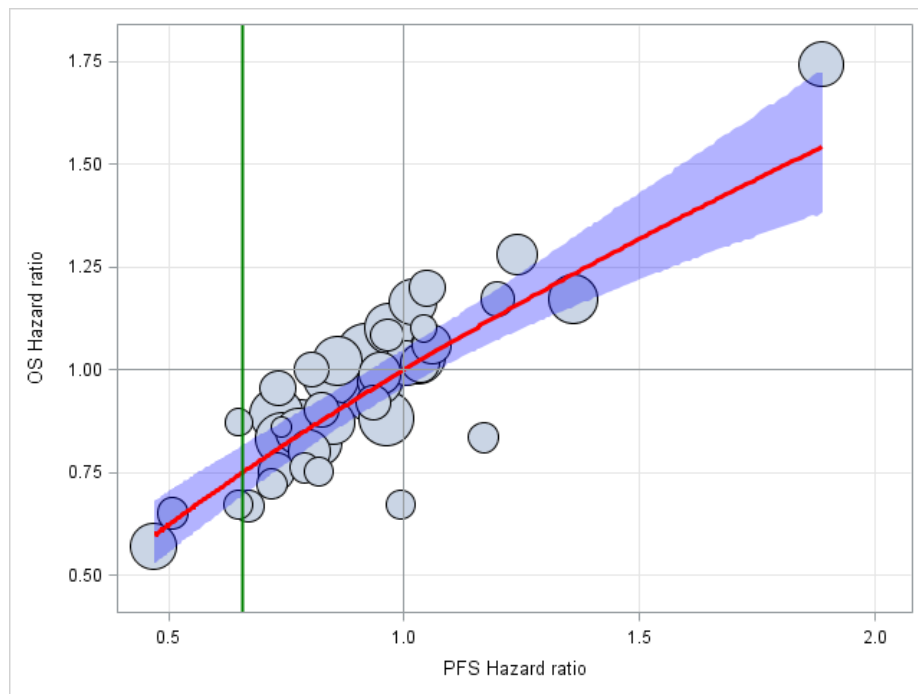
- Prior

- $\beta \sim N(0, 100^2), \tau \sim Uniform(0, 2)$

- Posterior*

- $\beta \sim N(0.6831, 0.0807^2)**$, $E(\tau) = 0.0241$

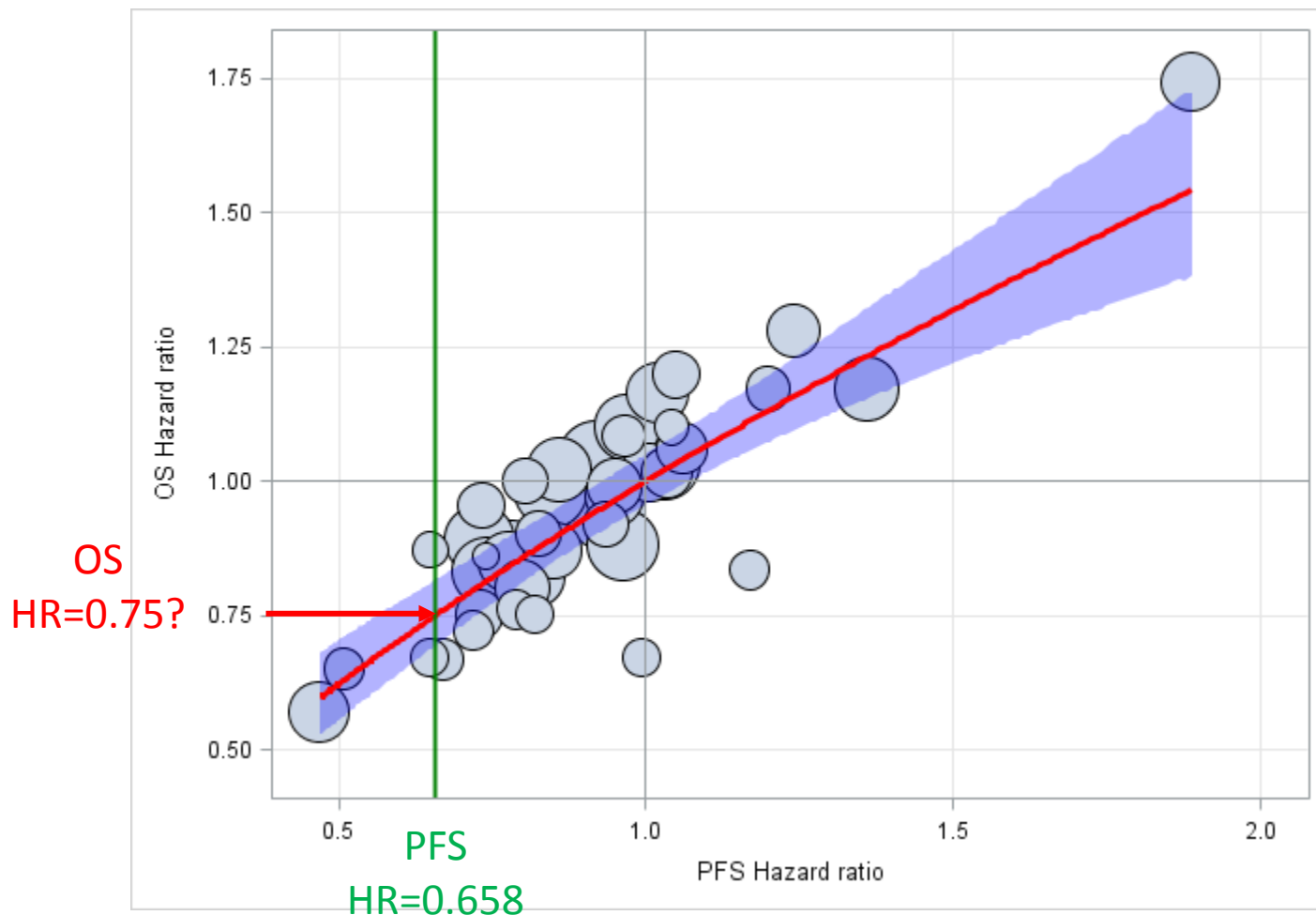
PFSとOSの関係性
(posteriorを用いて回帰直線+95%CI)



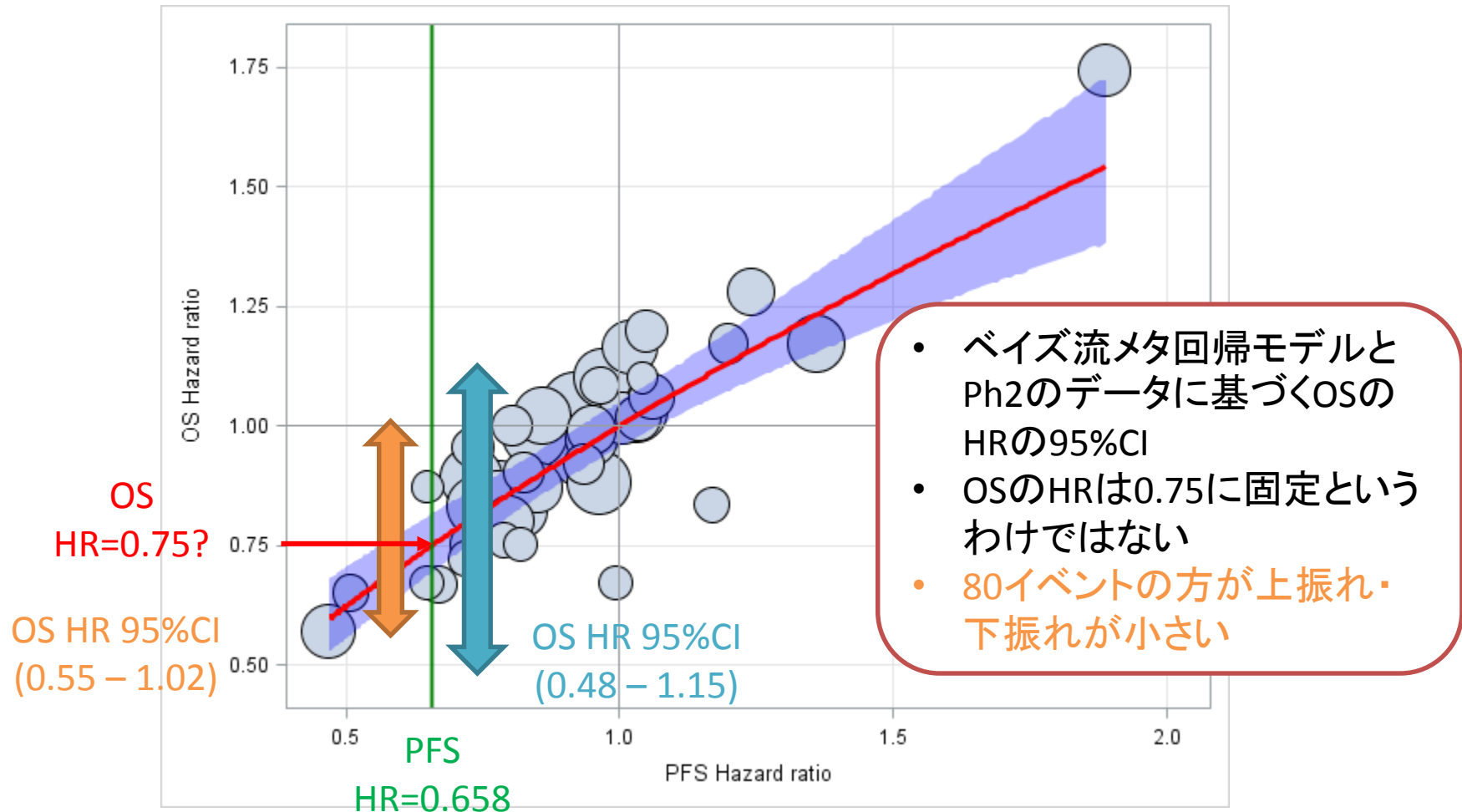
*Sabin et al. (2014)のsupplementより引用したが本文の図と異なるため、結果も多少異なる

**事後分布も正規分布と仮定(この後の解析のため)

OS HR=0.75としてPh3試験を計画するとしても、
80イベントのPh2と40イベントのPh2で違いはある？



例えば、80イベントのPh2と40イベントのPh2で、
想定できるOSのばらつきは違う(とも考えられる)



ベイズ流メタ回帰で得られたモデルと Ph2のPFS HRを用いたOSの95%CI

モデル式

– $\hat{\theta} \sim N(\mu, \epsilon^2)$

– $\mu \sim N(\phi, \tau^2)$

– $\phi = \beta \times \delta$

– $d \sim N(\delta, \sigma_d^2)$

– δ : Ph2の真のPFSのln HR

– d : Ph2で観測されたPFSのln HR

– σ_d : 上記の標準誤差

– Prior (メタ回帰の結果を事前情報に)

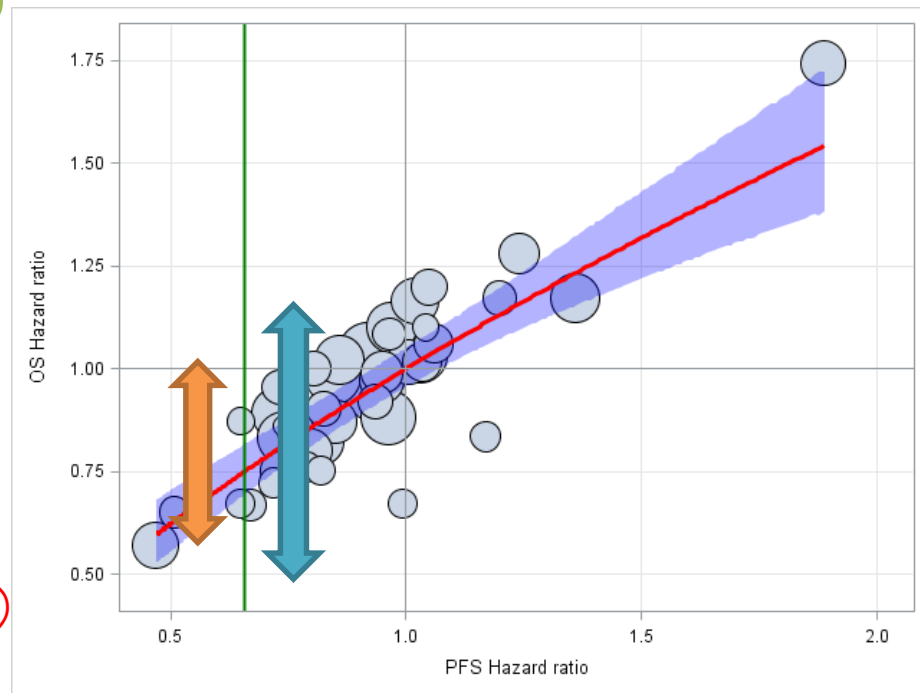
– $\beta \sim N(0.6831, 0.0807^2)$

– $\tau = 0.0241$

– $\delta \sim N(0, 10^2)$

緑: data
赤: parameter

PFSとOSの関係性

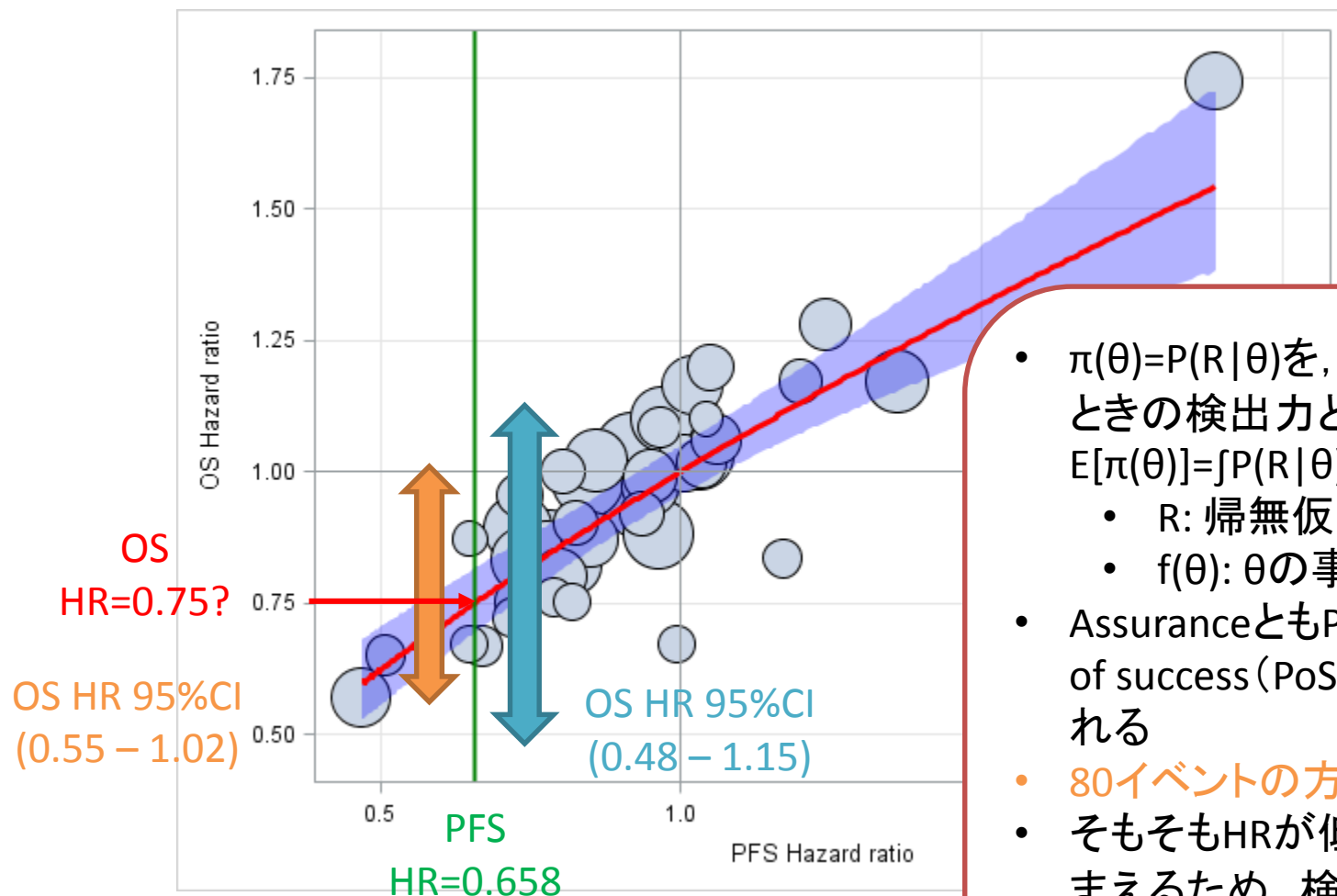


– $PFS\ HR = \exp(d) = 0.658, \sigma_d^2 = 0.05$

– Posterior mean, 95%CI

– $\mu = 0.75 (0.55, 1.02)$

OS HR=0.75 (検出力90%)としたPh3試験の計画に対し、 OSの事後分布で検出力の期待値を取ること



- $\pi(\theta)=P(R|\theta)$ を, $HR=\theta$ としたときの検出力とすると,
 $E[\pi(\theta)]=\int P(R|\theta)f(\theta) d\theta$
 - R: 帰無仮説の棄却
 - f(θ): θ の事後分布
- AssuranceともProbability of success (PoS)とも呼ばれる
- 80イベントの方が高そう
- そもそもHRが低い値も踏まえるため, 検出力より値は低くなる

PoSを実際に計算してみる

上: PFS HR=0.658だった場合のPh2のイベント数別のPoS

Observed PFS HR=0.658			Ph2 events		
True OS HR	Power	Ph3 events	40 (OS 95%CI: 0.48-1.15)	80 (OS 95%CI: 0.55-1.02)	120 (OS 95%CI: 0.57-0.96)
0.75	90%	508	67.8%	72.3%	75.9%
	80%	380	63.5%	66.2%	69.4%
Observed PFS HR=0.750			Ph2 events		
True OS HR	Power	Ph3 events	40 (OS 95%CI: 0.53-1.27)	80 (OS 95%CI: 0.60-1.11)	120 (OS 95%CI: 0.64-1.05)
0.75?? (0.82)	90%	508	53.6%	55.8%	55.1%
	80%	380	48.7%	49.5%	48.3%

- 予想したように、PFS HR=0.658の場合、イベント数が増えれば、PoSは上昇
- 検出力が高い試験デザインの方が、PoSも高い(平均的な検出力なので)

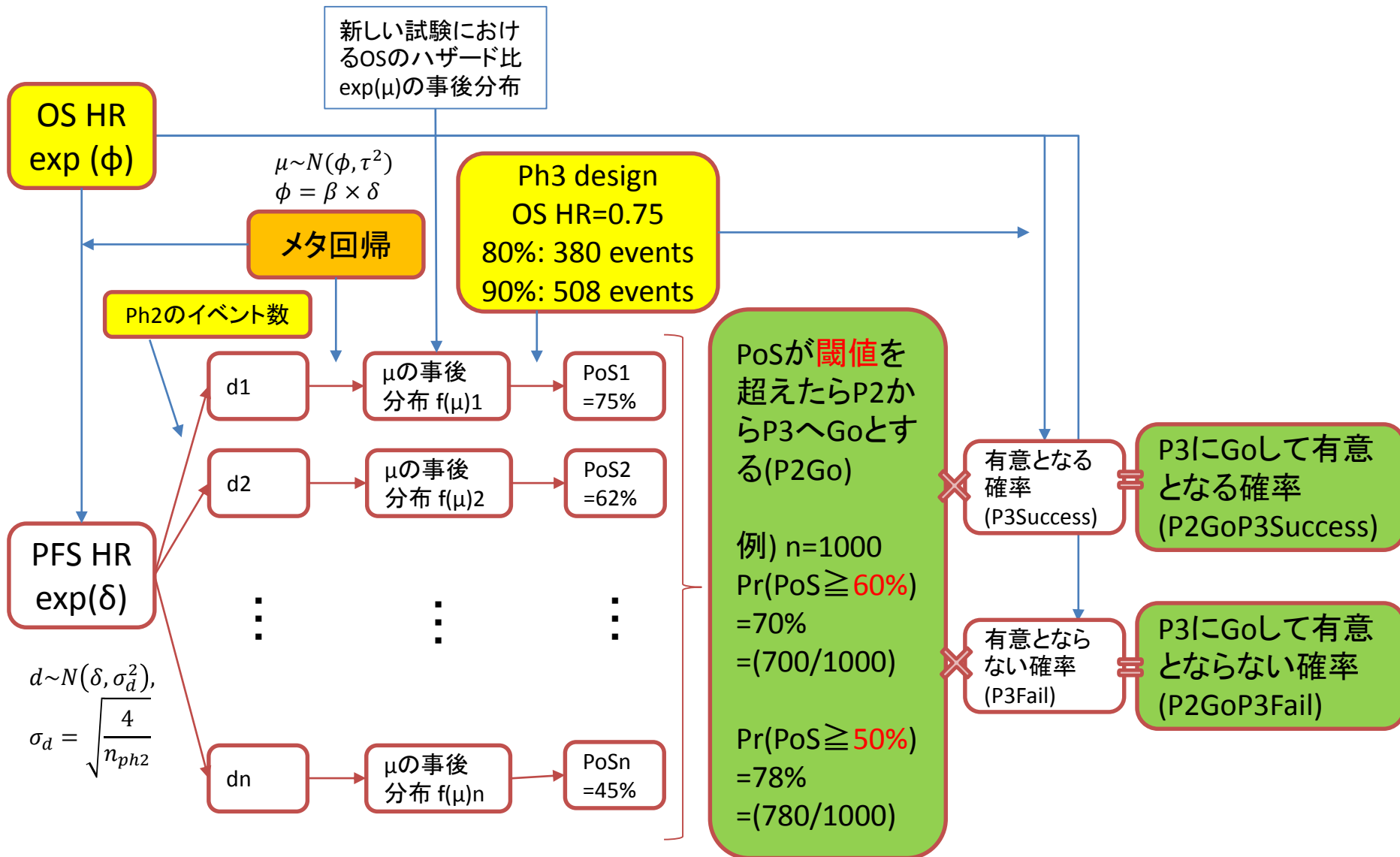
PoSを実際に計算してみる

下: 仮にPFSの結果を過大評価し, OSの治療効果も同程度と仮定

Observed PFS HR=0.658			Ph2 events		
True OS HR	Power	Ph3 events	40 (OS 95%CI: 0.48-1.15)	80 (OS 95%CI: 0.55-1.02)	120 (OS 95%CI: 0.57-0.96)
0.75	90%	508	67.8%	72.3%	75.9%
	80%	380	63.5%	66.2%	69.4%
Observed PFS HR=0.750			Ph2 events		
True OS HR	Power	Ph3 events	40 (OS 95%CI: 0.53-1.27)	80 (OS 95%CI: 0.60-1.11)	120 (OS 95%CI: 0.64-1.05)
0.75?? (0.82)	90%	508	53.6%	55.8%	55.1%
	80%	380	48.7%	49.5%	48.3%

- 予想したように, PFS HR=0.658の場合, イベント数が増えれば, PoSは上昇
- 検出力が高い試験デザインの方が, PoSも高い(平均的な検出力なので)
- 無理やりOS HR=0.75を仮定すれば検出力90%の試験とすることができるが, PoSを計算すると, OSの過大評価分, 低い値が算出される
- でも, このPoSって, 何に使えるの? → P3へのGo/NoGoの閾値に使える

PoSをGo/NoGoに用いる場合の性能評価のイメージ



数値実験に基づくPoSの閾値設定

			Range of true OS HR		
			0.75	0.4 to 1.0	0.85 to 1.0
PoS criterion	Ph2 events	Power	P(P2Go)	Max P(P2GoP3Fail)	Average P(P2GoP3Fail)
60%	80	90%	<u>70%</u>	19%	<u>13%</u>
		80%	62%	18%	10%
50%	80	90%	78%	26%	19%
		80%	73%	26%	16%
60%	40	90%	55%	20%	18%
		80%	49%	20%	15%
50%	40	90%	<u>70%</u>	28%	<u>26%</u>
		80%	62%	27%	23%

- 【Ph2 80イベント】PoS=60%を閾値にすると、OSで効果がある場合 (HR=0.75) にGoし、効果がない場合 (HR=0.85-1.0) にあまりGoして失敗しない → いい基準かも
- 【Ph2 40イベント】PoS=50%を閾値にすると、OSで効果がある場合 (HR=0.75) にGoするが、効果がない場合 (HR=0.85-1.0) でも、**26%でGoして失敗する**

ディスカッションポイント

- Ph2の結果からPh3の治療効果を見極める方法(今回はベイズ流メタ回帰)は有用か？
- 治療効果を固定した検出力に対して, 治療効果のばらつきを踏まえて検出力の期待値をとったPoS (Probability of Success)*という考え方は有用か？

* なお, successは有意/非有意以外にも設定可能であるため, 検出力の期待値をとったものだけがPoSというわけではない

Reference

- Sabin, T., Matcham, J., Bray, S., Copas, A., & Parmar, M. K. B. (2014). A Quantitative Process for Enhancing End of Phase 2 Decisions. *Statistics in Biopharmaceutical Research*, 6(1), 67–77.
<http://doi.org/10.1080/19466315.2013.852617>
- Sabin, T., Matcham, J., Copas, A., & Parmar, M. K. B. (2015). Assessing End of Phase 2 Decision Criteria. *Statistics in Biopharmaceutical Research*, 7(3), 230–237.
<http://doi.org/10.1080/19466315.2015.1064467>