Statistical assessment of consistency of treatment effect in multiregional clinical trials

Kevin J Carroll

Content and flow

- Background and key questions
- Design MRCT
- Analysis MRCT
- Examples
- Summary

Background

- Modern drug development is impossible without the use of MRCTs
- However, there are challenges as reflected in ICH E5 relating to possible inconsistency of effects from region to region
  - 'intrinsic' = race, biology, genetics
  - 'extrinsic' = cultural, social, economic, medical practice, quality of care, experience in clinical trials, quality of trial conduct and monitoring
- With more recent regulatory guidelines from PMDA and EMA together with notable test cases, MRCTs subject to a lot of attention within and outside of statistics

Content and flow

- Background and key questions
- Design MRCT
- Analysis MRCT
- Examples
- Summary

Construct

- Trial designed to detect a true overall effect θ with α 1-sided Type I error a and power 1-β. The information content is therefore

$$V = \frac{\left(z_\alpha + z_\beta\right)^2}{\theta^2}$$

- i=1 to r regions each with a fraction f_i patients. The estimated treatment effect estimate in i^{th} region is therefore

$$\hat{\theta}_i \sim N\left(\theta, \frac{1}{f_i V}\right) \text{ so that the overall estimate } \hat{\theta} \sim \sum_{i=1}^{r} f_i \hat{\theta}_i \sim N\left(\theta, \frac{1}{V}\right)$$

- However, if regional effects are considered as random then $\theta_i \sim N(\theta, \tau^2)$

$$\text{so that } \hat{\theta}_i \sim N\left(\theta, \tau^2 + \frac{1}{f_i V}\right) \text{ and } \hat{\theta} \sim \sum_{i=1}^{r} f_i \hat{\theta}_i \sim N\left(\theta, \tau^2 \sum_{i=1}^{r} f_i^2 + \frac{1}{V}\right)$$

Hung (2010)

- If, in truth, $\tau^2 > 0$, then Type I error will be inflated

$$1-\Phi\left(z_\alpha\left(1+V\tau^2\sum_{i=1}^{r}f_i^2\right)^{-0.5}\right)$$

- V should be increased to

$$\tilde{V}=\left(\frac{\theta^2}{(z_\alpha+z_\beta)^2}-\tau^2\sum_{i=1}^{r}f_i^2\right)^{-1} \quad \text{so that} \quad \frac{V}{\tilde{V}}=1-\frac{\tau^2}{\theta^2}(z_\alpha+z_\beta)^2\sum_{i=1}^{r}f_i^2$$

and power = $\Phi\left(-z_\alpha+(z_\alpha+z_\beta)\left(1+V\tau^2\sum_{i=1}^{r}f_i^2\right)^{-0.5}\right)$

- For $V=\frac{N}{\sigma^2}$ sample size inflation is:

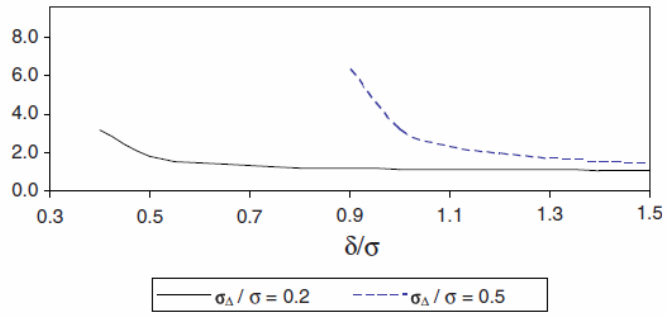
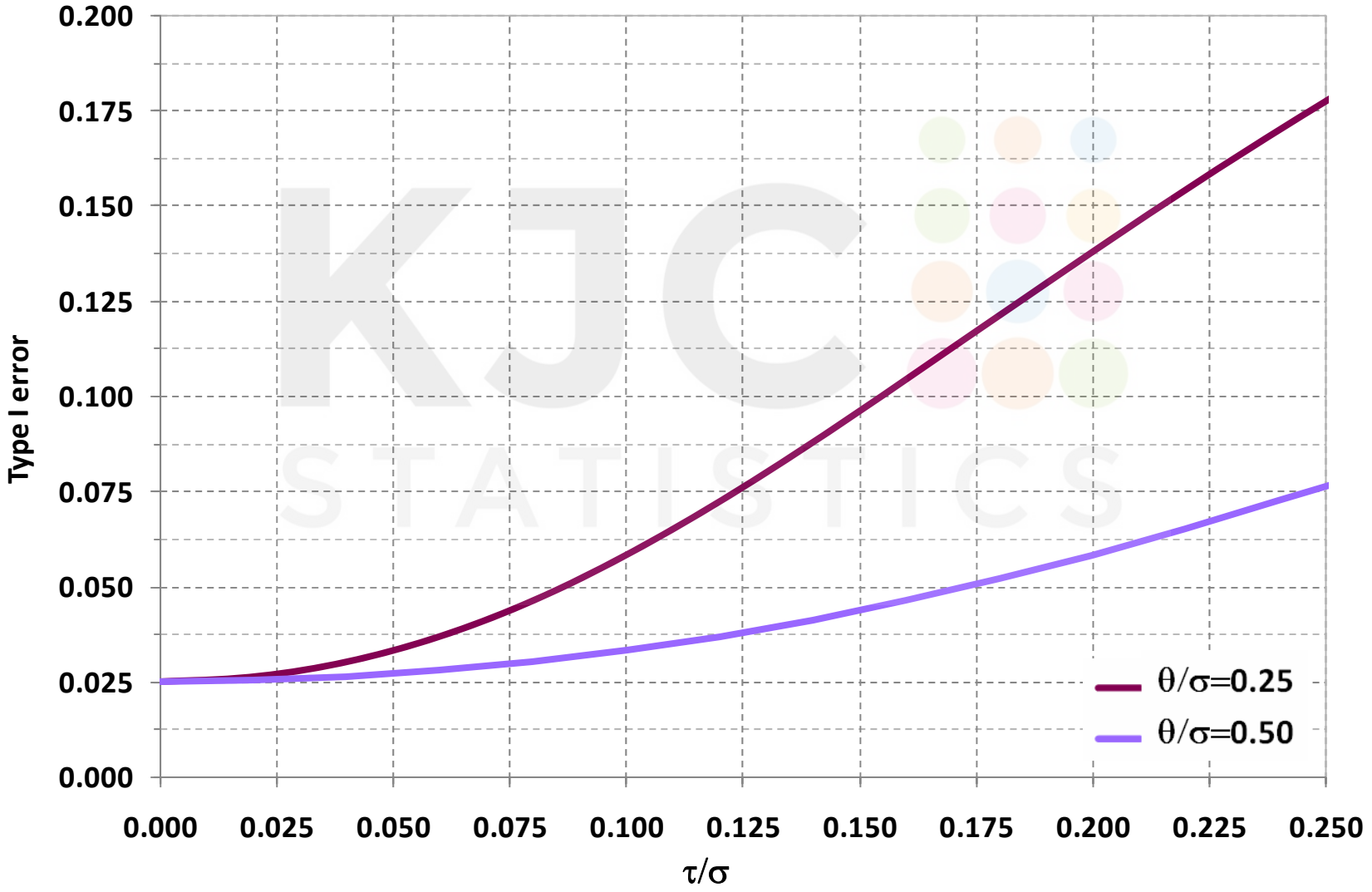


Figure 1. Sample size ratio N/N_0 versus (δ/σ).

Type I error inflation for $V=N/\sigma^2$, r=3, $f_i=1/3$, $\beta=0.1$



Issues and Questions (1)

- Ignoring τ^2 increases Type I error and reduces power.
- Should N in an MRCT be routinely increased, allowing for τ^2?
- In trial planning, what value for τ^2 (or τ / σ) should be assumed? On what is the choice based? Is it logical to assume $\tau^2 > 0$?
- At what value of τ / σ does a MRCT become
  - infeasible in terms of trial size
  - impossible to interpret clinically due to excessive between region variation in treatment effect?

Issues and Questions (2)

- How do we define ‘region’? Should there be a regulatory standard agreed to cover all trials ?
- What should be the allocation of N across regions / countries? How do we determine This?
- What is meant by ‘consistency’ ? How do we define this? How do we assess it? What is the value and role of routine homogeneity testing of regional results? And graphical methods?
- Should a random effects analysis be the standard in MRCTs? What are the consequences if so?

What allocation of patient to regions?

- If we want to observe a positive treatment effect in each region, then

$$\Pr\left(\hat{\theta}_i > 0, \forall\, i = 1 \text{ to } r\right) = \prod_{i=1}^{r} \Phi\left\{\sqrt{f_i}\left(z_\alpha + z_\beta\right)\right\}$$

- But need also to observe an overall significant effect, so then we want

$$\Pr\left(\hat{\theta}_i > 0, \forall\, i = 1 \text{ to } r \cap \hat{\theta} > z_\alpha V^{-0.5}\right)$$

- This can be derived noting that where

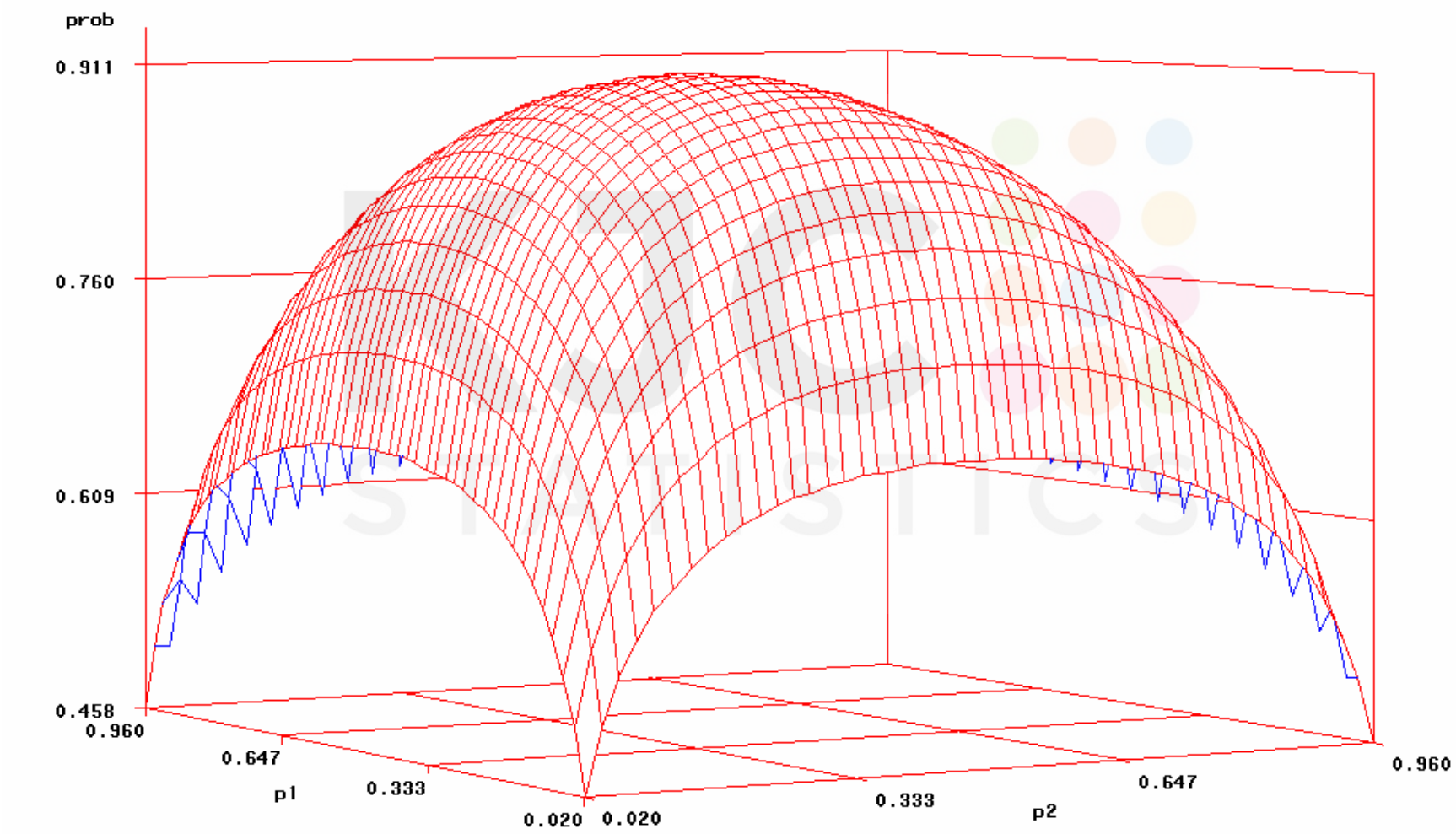
$$\left(\hat{\theta}, \hat{\theta}_1, \ldots, \hat{\theta}_r\right) \sim N\left(\underline{\theta}, \underline{\Lambda}\right)$$

$$\underline{\Lambda} = \begin{bmatrix} \frac{1}{V} & \frac{1}{V} & . & \frac{1}{V} \\ \frac{1}{V} & \frac{1}{f_1 V} & 0 & 0 \\ . & 0 & . & 0 \\ \frac{1}{V} & 0 & 0 & \frac{1}{f_r V} \end{bmatrix}$$

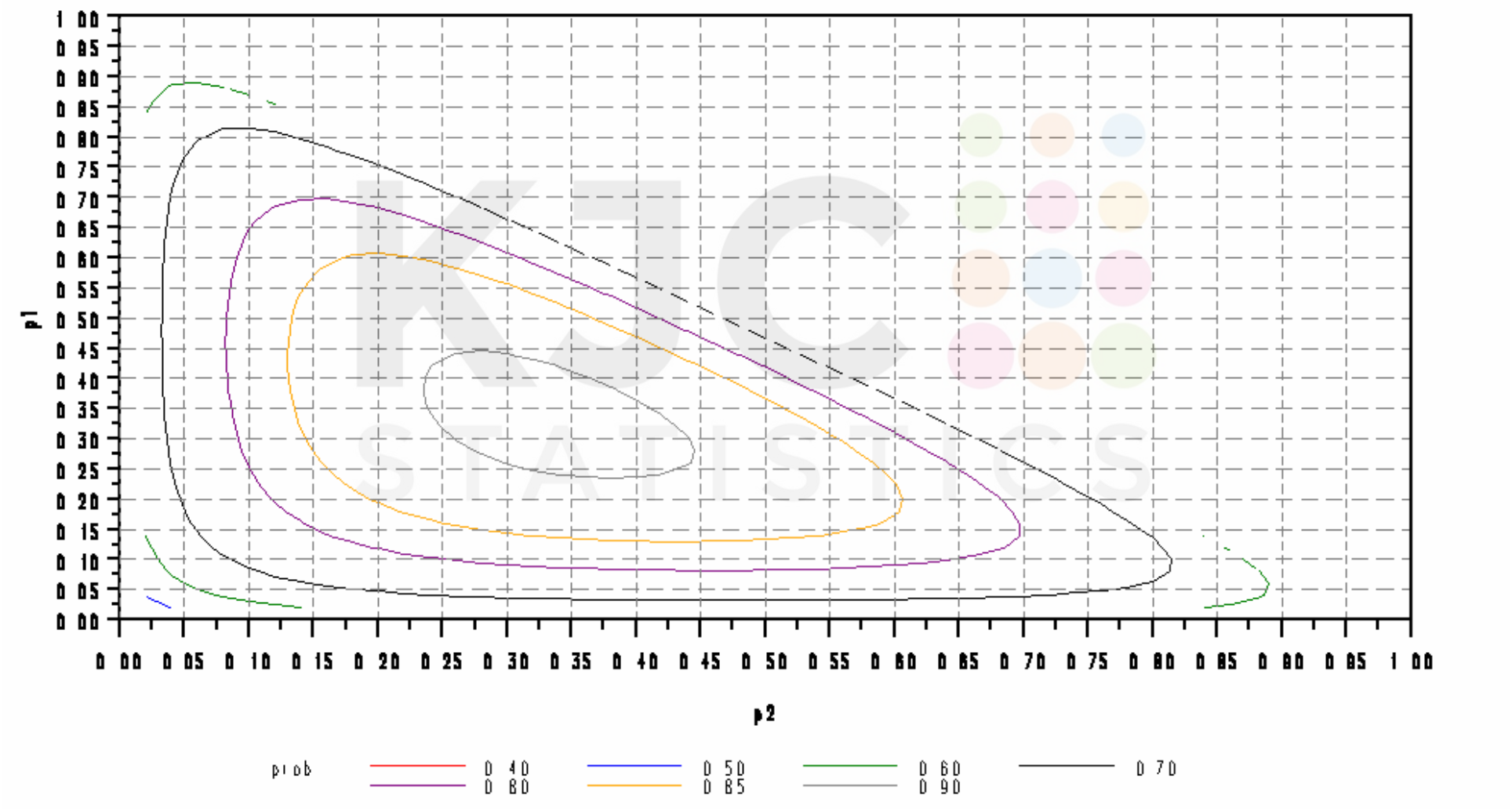
- Chuang (2008) considers the most typical MRCT with 3 regions, to reflect US, EU and Japan concluding the smallest region could be as low as 15% and yet still achieve = 0.80.

$$\Pr\left(\hat{\theta}_i > 0, \forall\, i = 1 \text{ to } r \cap \hat{\theta} > z_\alpha V^{-0.5}\right)$$

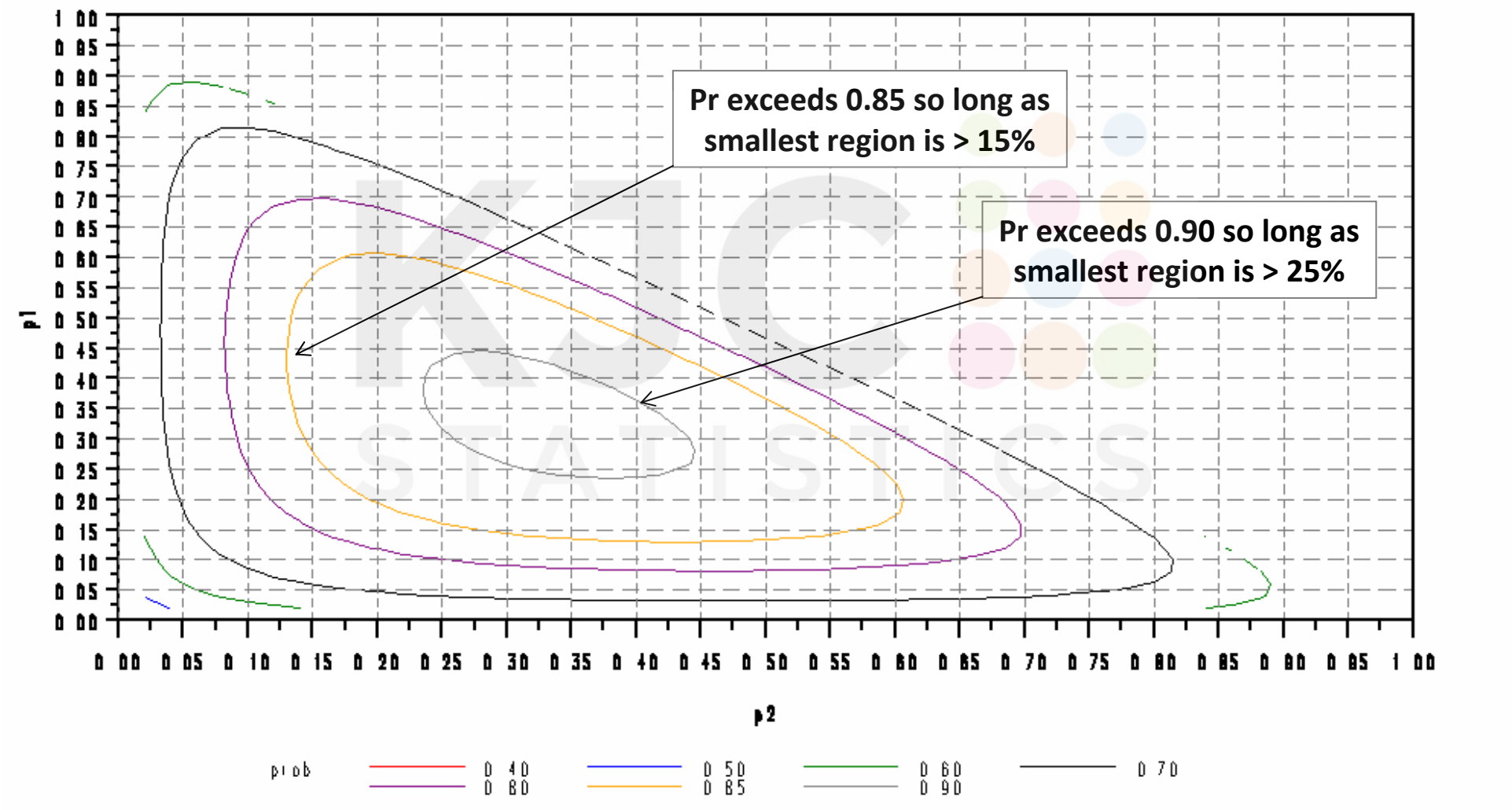
Pr observing a positive treatment effect in each region



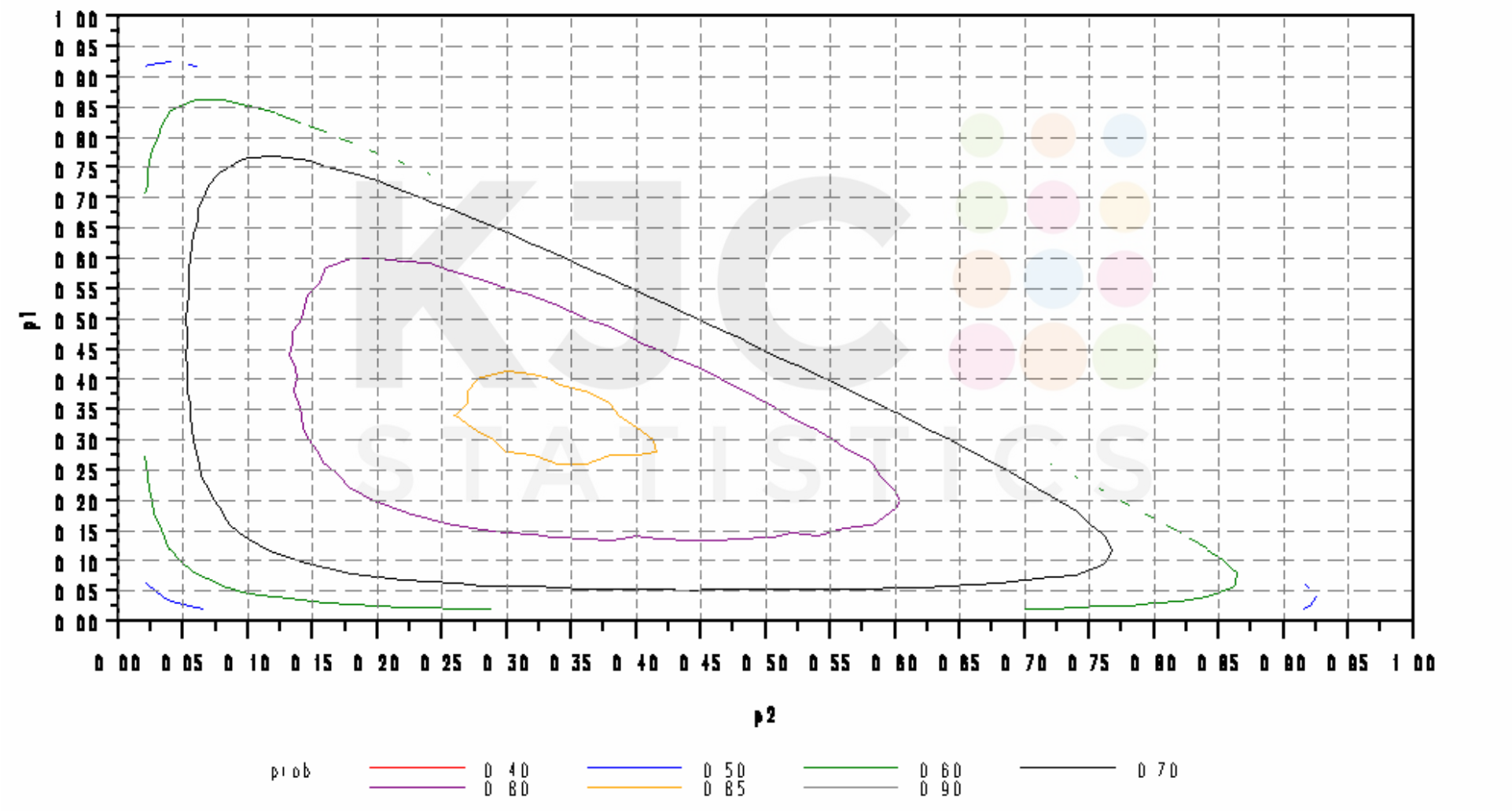
Pr observing a positive treatment effect in each



Pr observing a positive treatment effect in each

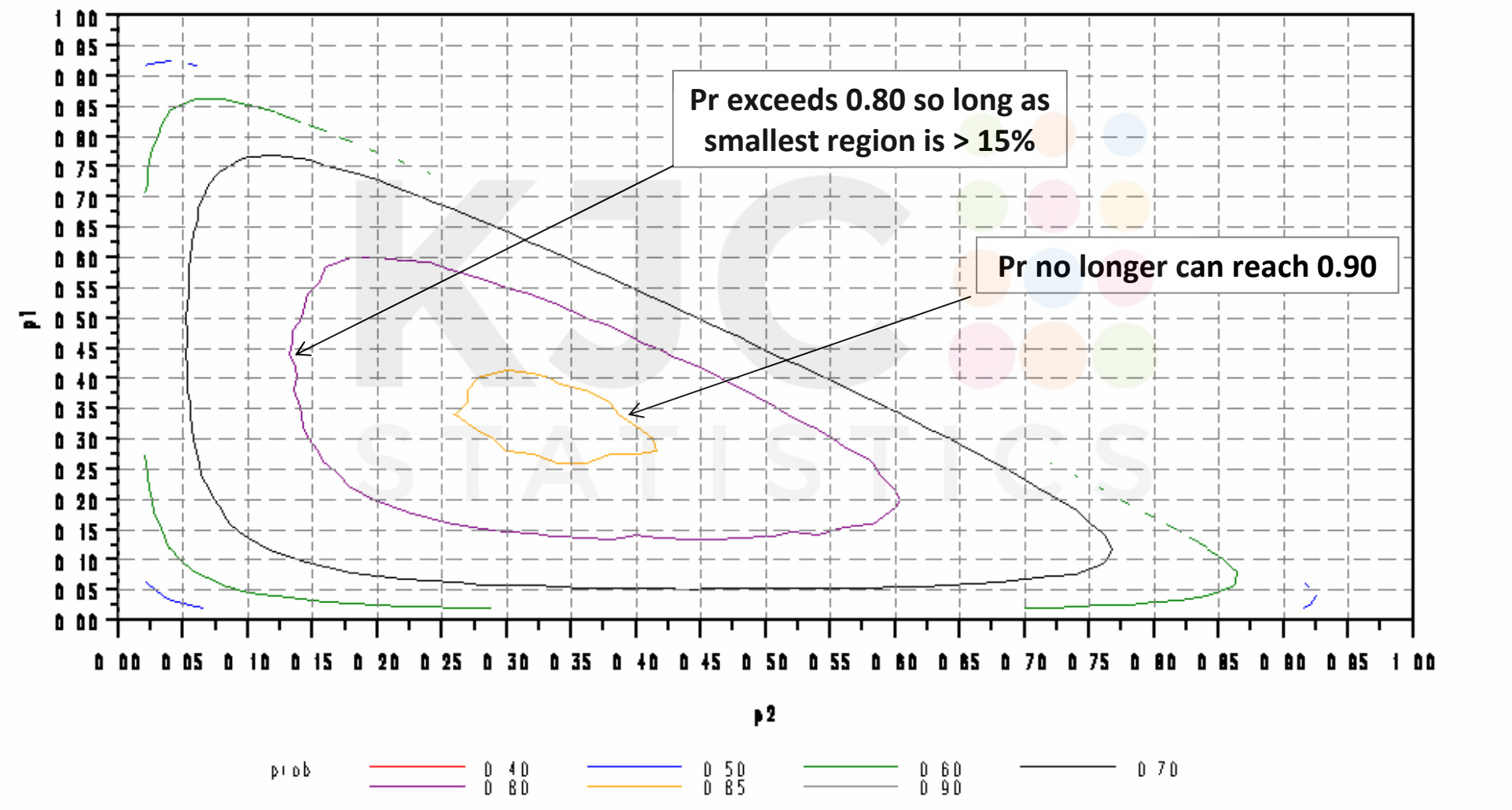


Pr observing a positive treatment effect in each region and p<0.05 for the overall effect



Based on 10,000 simulations for each combination of p1 and p2

Pr observing a positive treatment effect in each region and p<0.05 for the overall effect



Based on 10,000 simulations for each combination of p1 and p2

Looking at the 'reference' region of interest[1]

- Denote $\hat{\theta}_{ref} \sim N\left(\theta, \frac{1}{fV}\right)$ and $\hat{\theta}_{nref} \sim N\left(\theta, \frac{1}{(1-f)V}\right)$ then consistency between the reference region and non-reference regions may be defined in several ways:

1. $\hat{\theta}_{ref} - 0.5\hat{\theta} > 0$
2. $\hat{\theta}_{ref} - 0.5\hat{\theta}_{nref} > 0$
3. $\left(\hat{\theta}_{ref} - \hat{\theta}_{nref}\right) + \frac{\theta}{2} > 0$
4. $\hat{\theta}_{ref} - \left(\hat{\theta} - z_{\alpha} V^{-0.5}\right) > 0$
5. $\hat{\theta}_{ref} > 0$ and $\hat{\theta}_{nref} > 0$

- Denoting criteria as C_j then pr(meeting criteria and achieving $\hat{\theta} > z_{\alpha} V^{-0.5}$ can be determined by noting

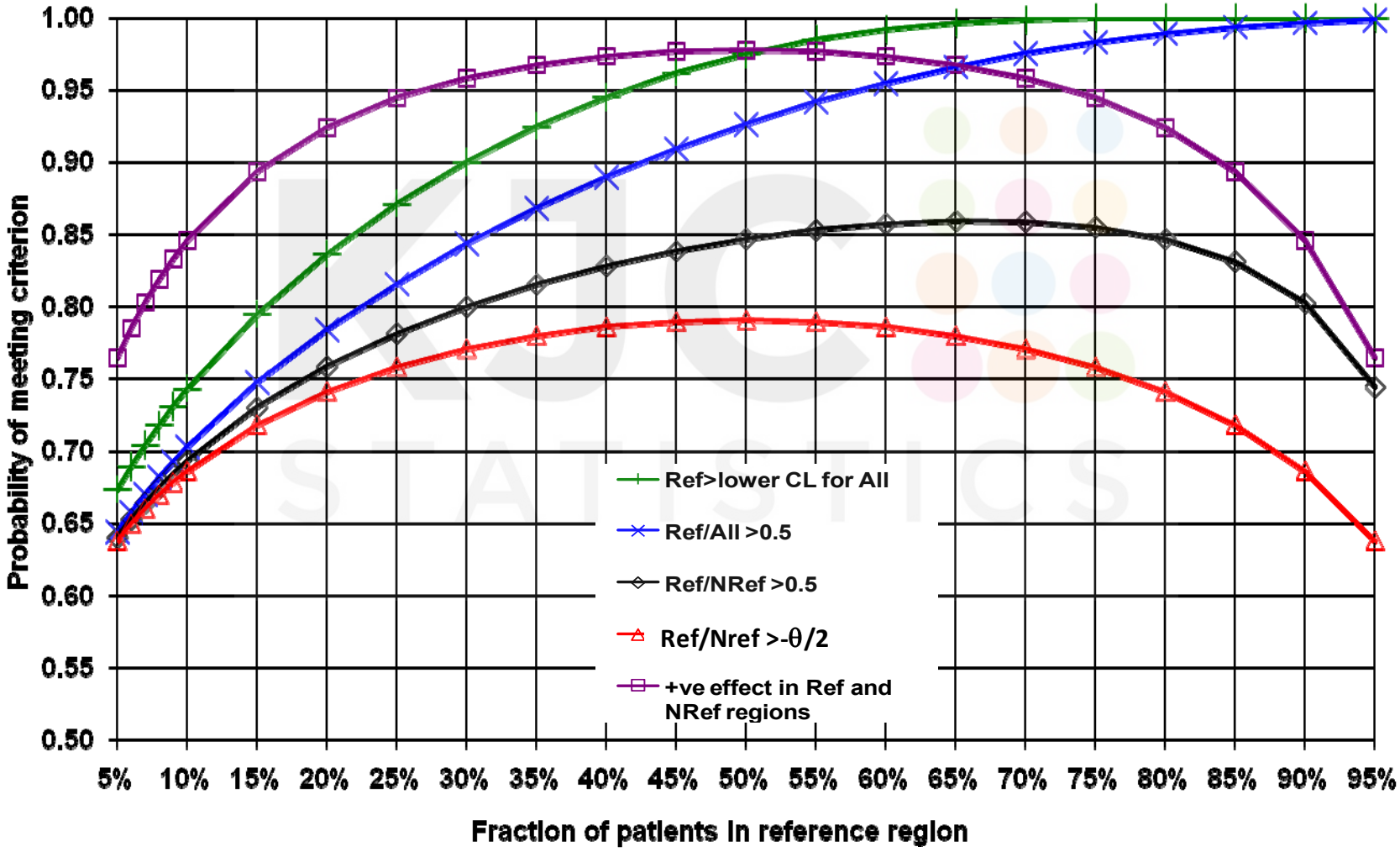
$$\begin{pmatrix} C_1 \\ \hat{\theta} \end{pmatrix} \sim N\left(\begin{matrix} \theta(1-k) \\ \theta \end{matrix}, \begin{matrix} \frac{1}{fV} + \frac{k(k-2)}{V} & \frac{1-k}{V} \\ \frac{1-k}{V} & \frac{1}{V} \end{matrix} \right)$$

$$\begin{pmatrix} C_2 \\ \hat{\theta} \end{pmatrix} \sim N\left(\begin{matrix} \theta(1-k) \\ \theta \end{matrix}, \begin{matrix} \frac{1}{fV} + \frac{k^2}{V(1-f)} & \frac{1-k}{V} \\ \frac{1-k}{V} & \frac{1}{V} \end{matrix} \right)$$

$$\begin{pmatrix} C_4 \\ \hat{\theta} \end{pmatrix} \sim N\left(\begin{matrix} z_{\alpha} V^{-0.5} \\ \theta \end{matrix}, \begin{matrix} \frac{(1-f)}{fV} & 0 \\ 0 & \frac{1}{V} \end{matrix} \right)$$

1: Carroll & LeMaulf (2011)

Pr(meeting criteria 1 to 5)



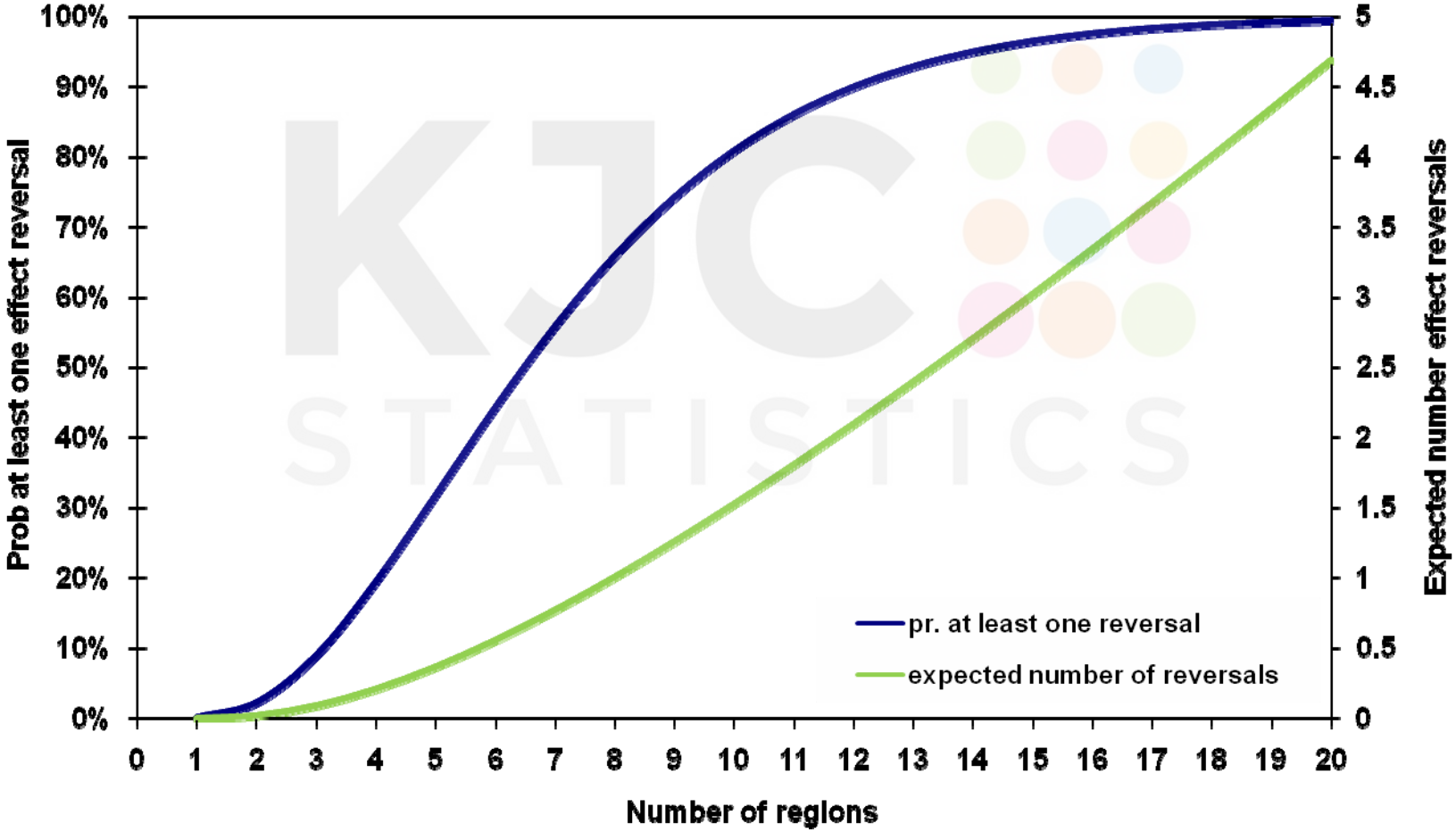
So what?

- Of these possible alternative criteria, only criterion #4 represents some improvement over the MHLW guideline criterion in terms of the fraction of reference region patients required in a MRCT
- Criterion #4 requires around 1/3 fewer patients than Criterion #1 for about the same overall power.

"Effect Reversals", where the treatment effect is positive overall but numerically negative in some regions are to be expected in a large multiregional trials

Senn, S and Harrell, F. On wisdom after the Event. *J Clin Epidemiol*, 1997. 50(7):749-751.

In a trial with 90% (80%) power and 7 (6) regions, the probability of at least one effect reversal >50%.



Senn, S and Harrell, F. On wisdom after the Event. *J Clin Epidemiol*, 1997. 50(7):749-751.

Content and flow

- Background and key questions
- Design MRCT
- Analysis MRCT
- Examples
- Summary

Random effects analysis – good idea?

- If regional effects are considered as random then $\theta_i \sim N(\theta, \tau^2)$ and

$$\hat{\theta}_i \sim N\left(\theta, \tau^2 + \frac{1}{f_i V}\right) \quad \text{with} \quad Q = V\sum_{i=1}^{r} f_i\left(\hat{\theta}_i - \hat{\theta}\right)^2 \sim \chi^2_{r-1}$$

- Then $\hat{\tau}^2 = \dfrac{Q-(r-1)}{1-\sum_{i=1}^{r} f_i^2}$ and $w_i = \left(\hat{\tau}^2 + \dfrac{1}{f_i V}\right)^{-1}$ so that

$$\hat{\theta}_{RE} \sim \sum_{i=1}^{r} w_i \hat{\theta}_i \sim N\left(\theta, \left(\sum_{i=1}^{r} w_i\right)^{-1}\right)$$

- Example, Chen (2010):

  $\hat{\theta}$ = 0.89 95% CI (0.79, 0.99), p=0.037

  $\hat{\theta}_{RE}$ = 0.91 (0.76, 1.08), p=0.29

Table I. PURSUIT: Efficacy by Region.

	N	Odds ratio	95% confidence interval
Overall	10948	0.89	(0.79, 0.99)
North America	4358	0.75	(0.63, 0.91)
Western Europe	4243	0.92	(0.77, 1.11)
Latin America	585	1.03	(0.60, 1.76)
Eastern Europe	1762	1.09	(0.85, 1.39)

N: sample size. Odds ratio: eptifibatide vs placebo, lower the better [19].

Random effects analysis – good idea?

- Example, Chen (2010):

  $\hat{\tau}^2 = 0.016$, $\hat{\tau}/\hat{\sigma} = 1.06$

  $\chi^2 = 6.3$, p=0.096, I^2=0.52

  $\hat{\theta} = 0.89$ CI (0.79, 0.99), p=0.037

  $\hat{\theta}_{RE} = 0.91$ CI (0.76, 1.08), p=0.29

Table I. PURSUIT: Efficacy by Region.

	N	Odds ratio	95% confidence interval
Overall	10948	0.89	(0.79, 0.99)
North America	4358	0.75	(0.63, 0.91)
Western Europe	4243	0.92	(0.77, 1.11)
Latin America	585	1.03	(0.60, 1.76)
Eastern Europe	1762	1.09	(0.85, 1.39)

N: sample size. Odds ratio: eptifibatide vs placebo, lower the better [19].

- $\hat{\theta}_{RE} = 0.91$ (0.68, 1.21), p=0.36 if Follmann[1] adjustment applied which means using a t value on k-1 df for $\hat{\theta}_{RE}$ as opposed to z value. This has been advocated in the recent past by FDA for meta-analyses .

- If applied this has some odd consequences for MRCTs.

	fi	RR	Lower 95% CL	Upper 95% CL	p-value
Region 1	0.33	0.80	0.65	0.99	0.0416
Region 2	0.33	0.80	0.65	0.99	0.0416
Region 3	0.33	0.80	0.65	0.99	0.0416
All		0.80	0.71	0.91	0.0004
RE		0.80	0.61	1.05	0.0718

1 Follmann (1999)

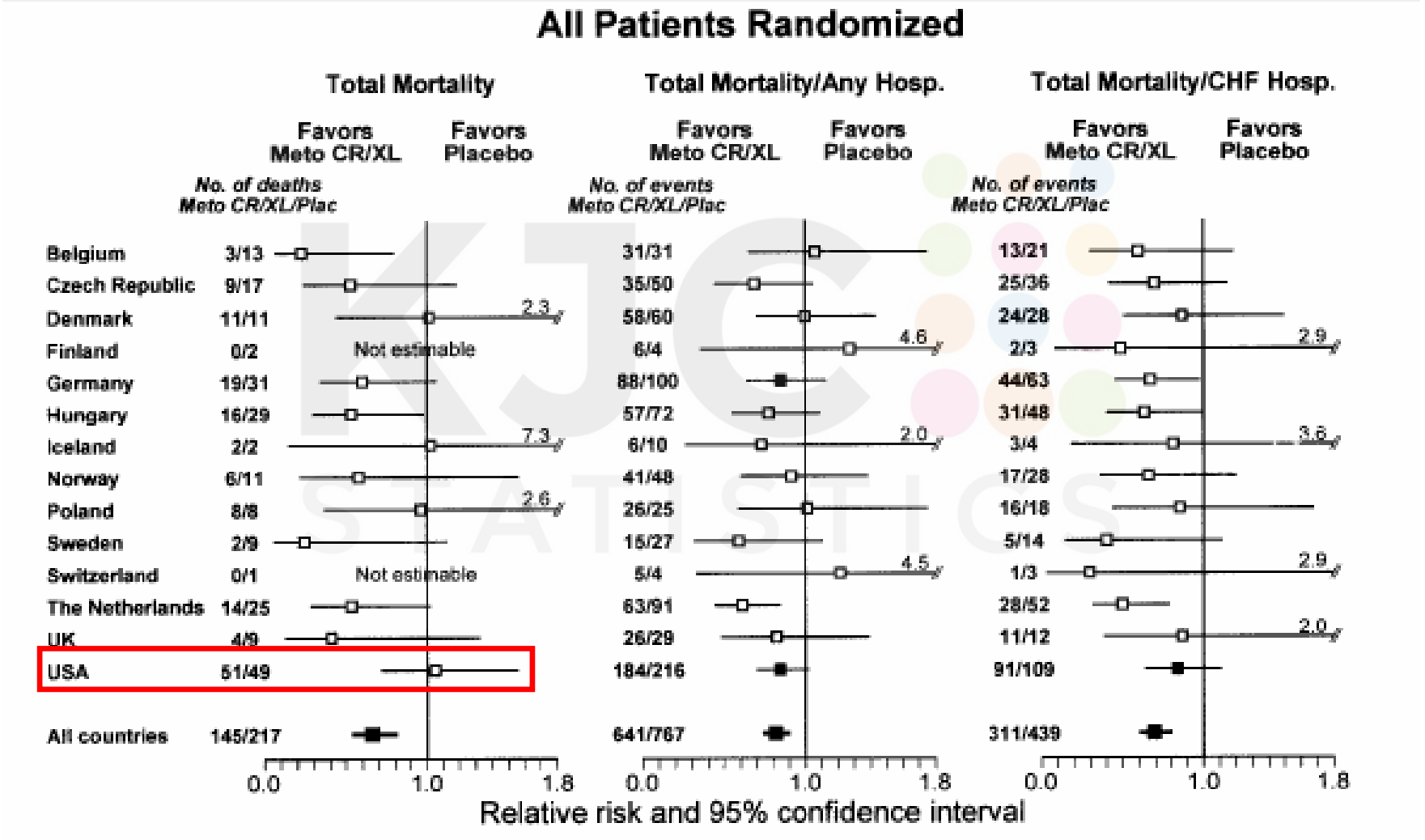
Content and flow

- Background and key questions
- Design MRCT
- Analysis MRCT
- Examples
- Summary

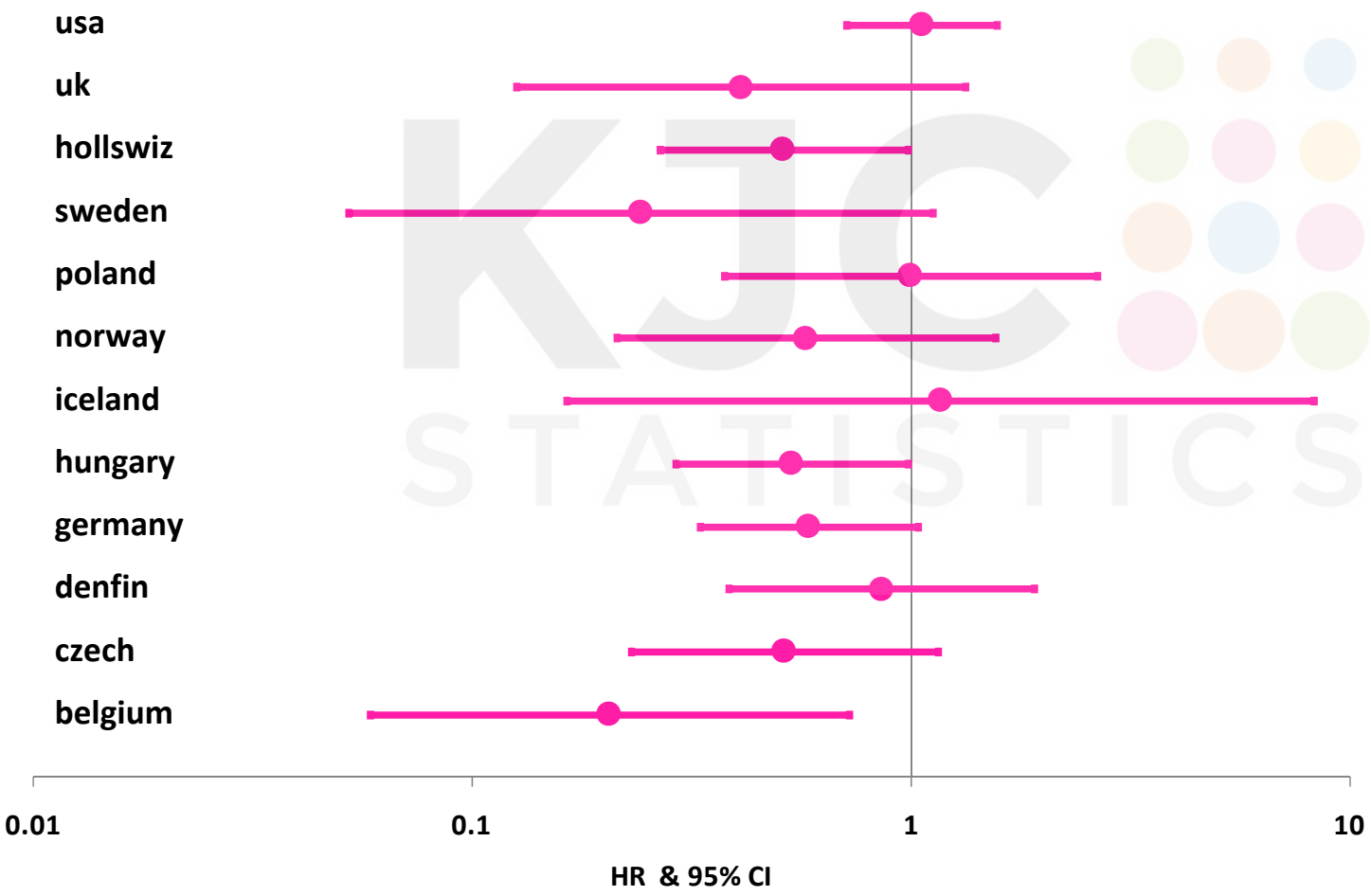
MERIT-HF

- Examined metoprolol CR/XL vs placebo in patients chronic heart failure and decreased ejection fraction
- Randomised 3991 across 14 countries
- The study stopped early on the recommendation of the IDMC. All-cause mortality was lower in the metoprolol CR/XL group than in the placebo group (145 [7·2%] vs 217 deaths [11·0 %]) RR = 0·66 [95% CI 0·53–0·81]; p=0·00009.
- However....

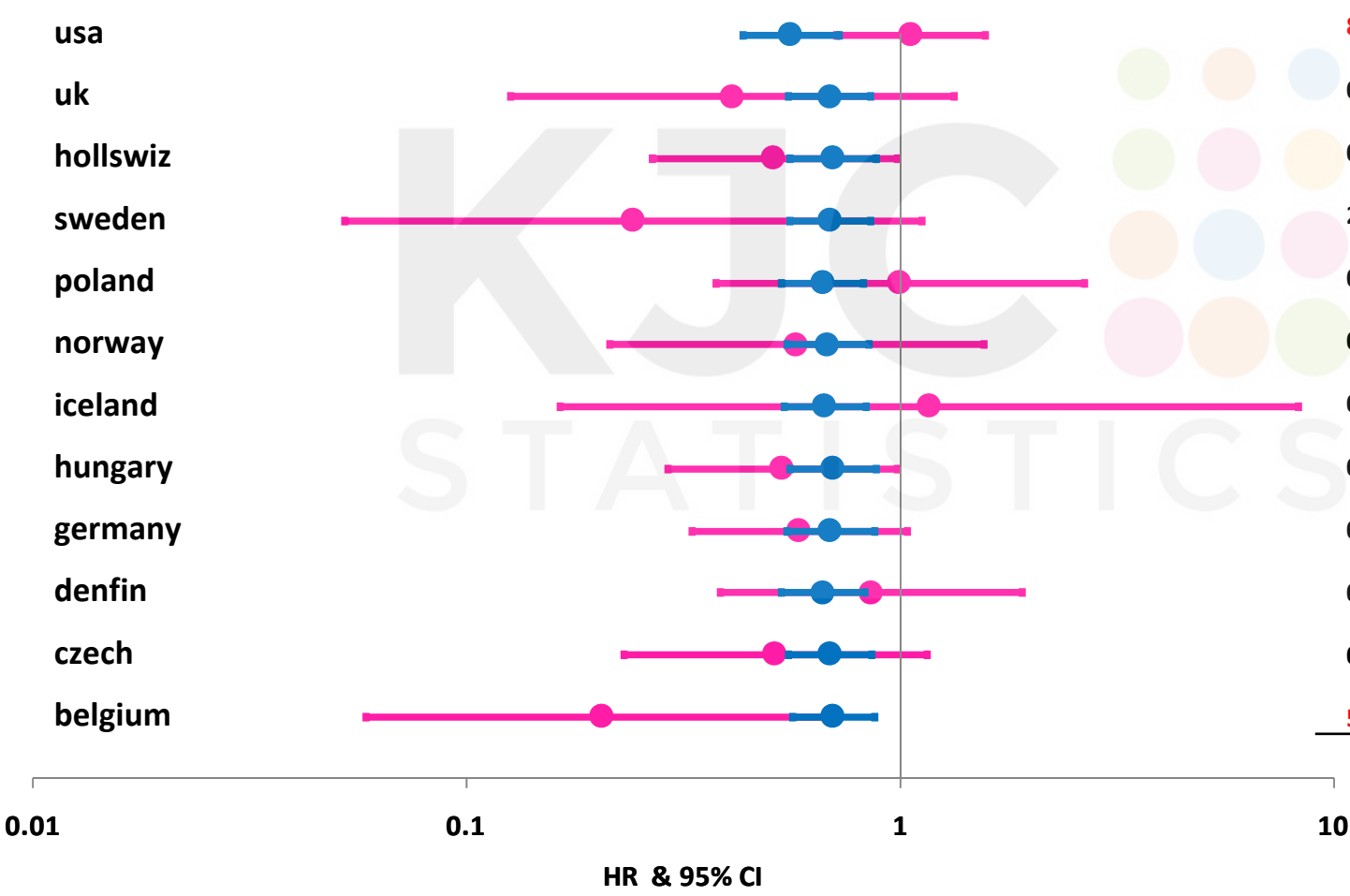
MERIT-HF: Inconsistent regional effects?



MERIT-HF: Mortality by region



MERIT-HF: Mortality by region vs all other regions

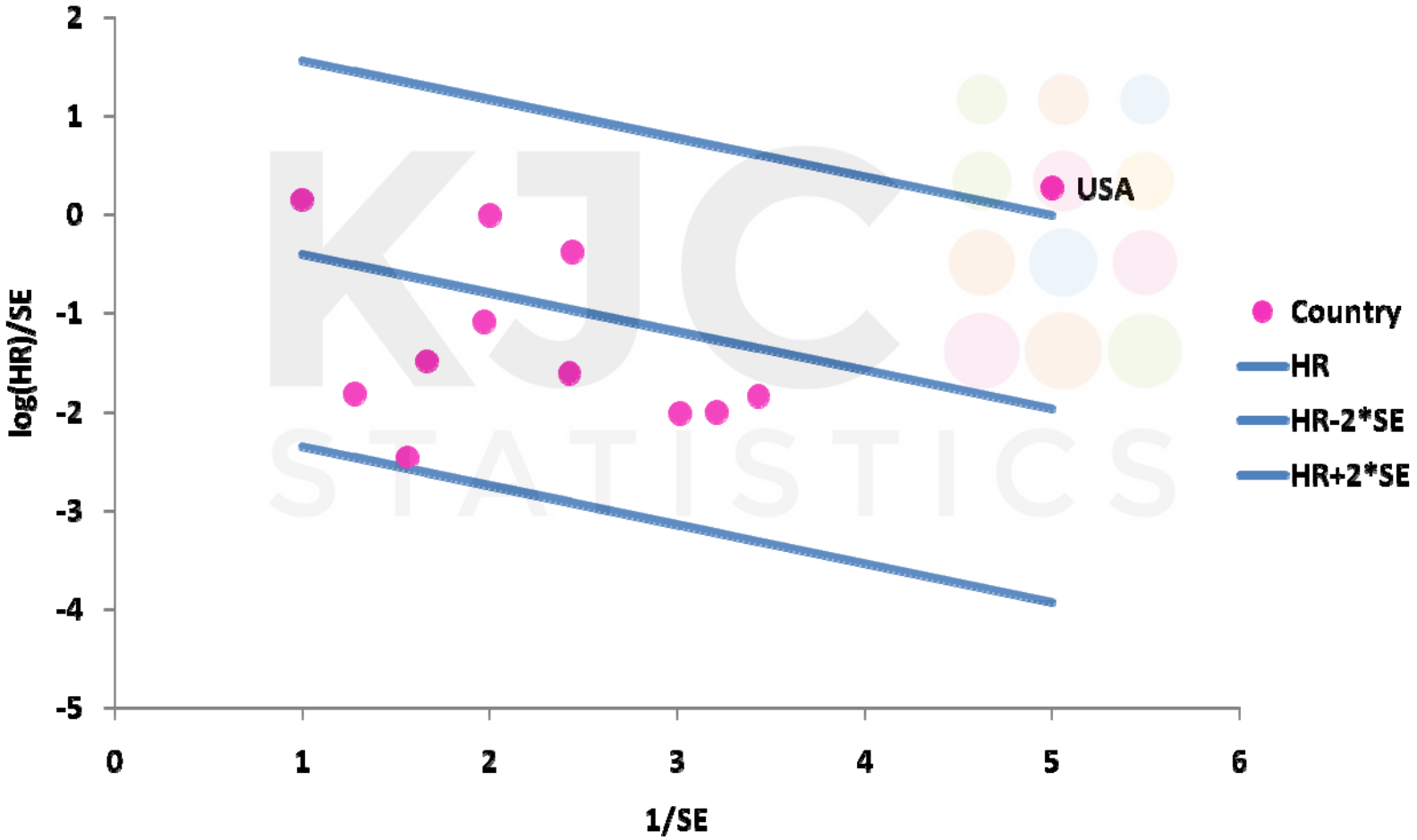


	t^2 for ● vs ●	f_i for ●	$t^2*(1-f_i)$	% total χ^2
usa	8.39	27.6%	6.07	34.4%
uk	0.70	3.6%	0.67	3.8%
hollswiz	0.60	11.0%	0.54	3.0%
sweden	2.74	3.0%	2.65	15.0%
poland	0.80	4.4%	0.76	4.3%
norway	0.06	4.7%	0.05	0.3%
iceland	0.35	1.1%	0.35	2.0%
hungary	0.45	12.4%	0.39	2.2%
germany	0.14	13.8%	0.12	0.7%
denfin	0.51	6.6%	0.48	2.7%
czech	0.36	7.2%	0.33	1.9%
belgium	5.48	4.4%	5.24	29.6%
			17.67	100.00%

χ^2 p-value =0.090

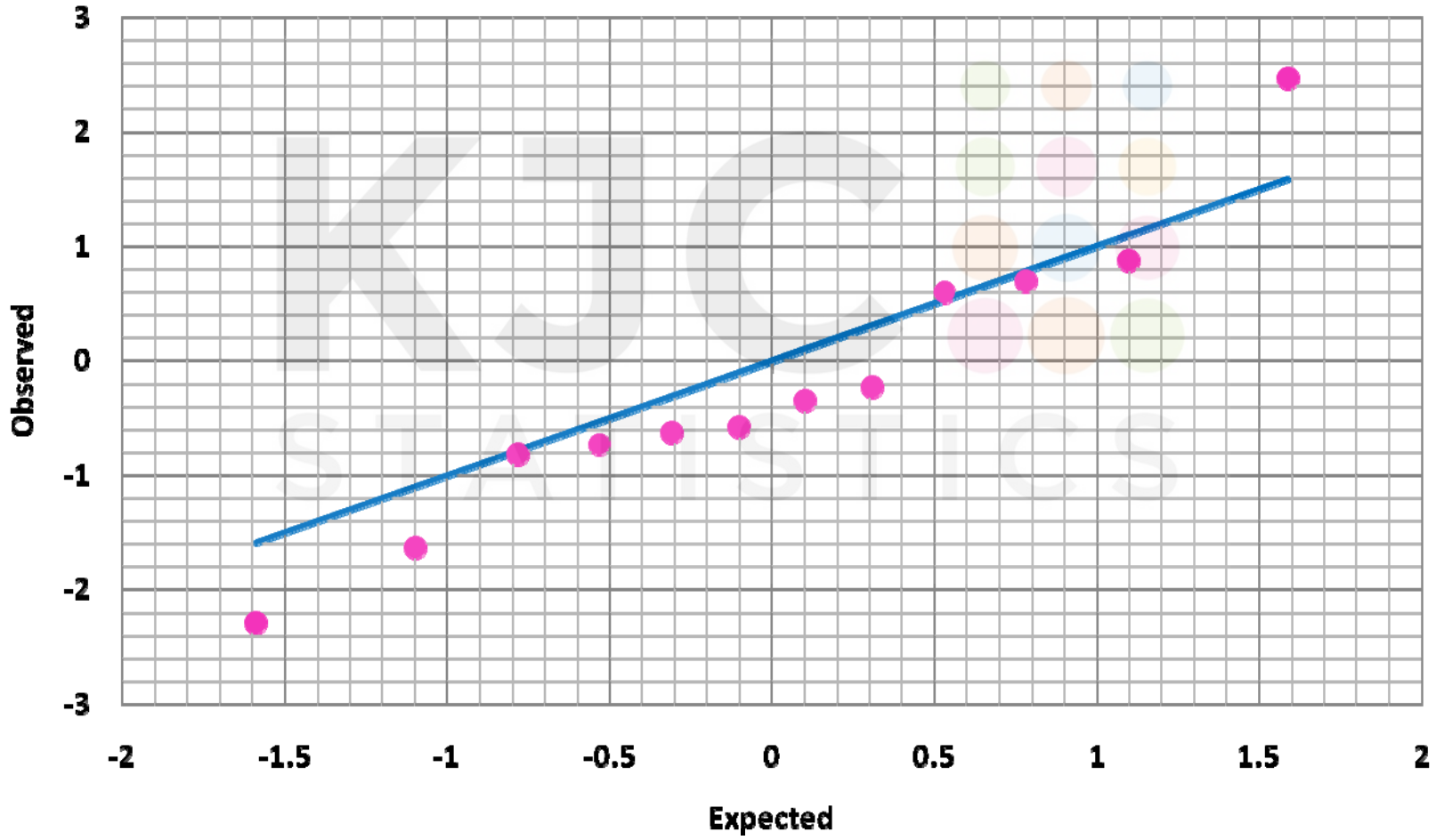
MERIT-HF: Mortality:

HR=0.65, 95% CI (0.49. 0.86), τ^2=0.038, s^2=0.14, I^2=0.21



MERIT-HF: Mortality:

HR=0.65, 95% CI (0.49. 0.86), τ^2=0.038, s^2=0.14, I^2=0.21



MERIT-HF: Mortality vs Primary

Mortality

	Drug		placebo			
Country	Events	N	Events	N	vi	xi
belgium	3	68	13	66	0.3804	-1.4962
czech	9	123	17	124	0.1537	-0.6279
Denm/Fin	11	161	13	164	0.1555	-0.1486
germany	19	252	31	247	0.0769	-0.5096
hungary	16	211	29	212	0.0875	-0.5900
iceland	2	19	2	22	0.9019	0.1466
norway	6	97	11	105	0.2377	-0.5269
poland	8	102	8	102	0.2304	0.0000
sweden	2	39	9	46	0.5637	-1.3390
holl/swiz	14	299	26	291	0.1031	-0.6462
uk	4	87	9	83	0.3376	-0.8580
usa	51	532	49	539	0.0363	0.0531

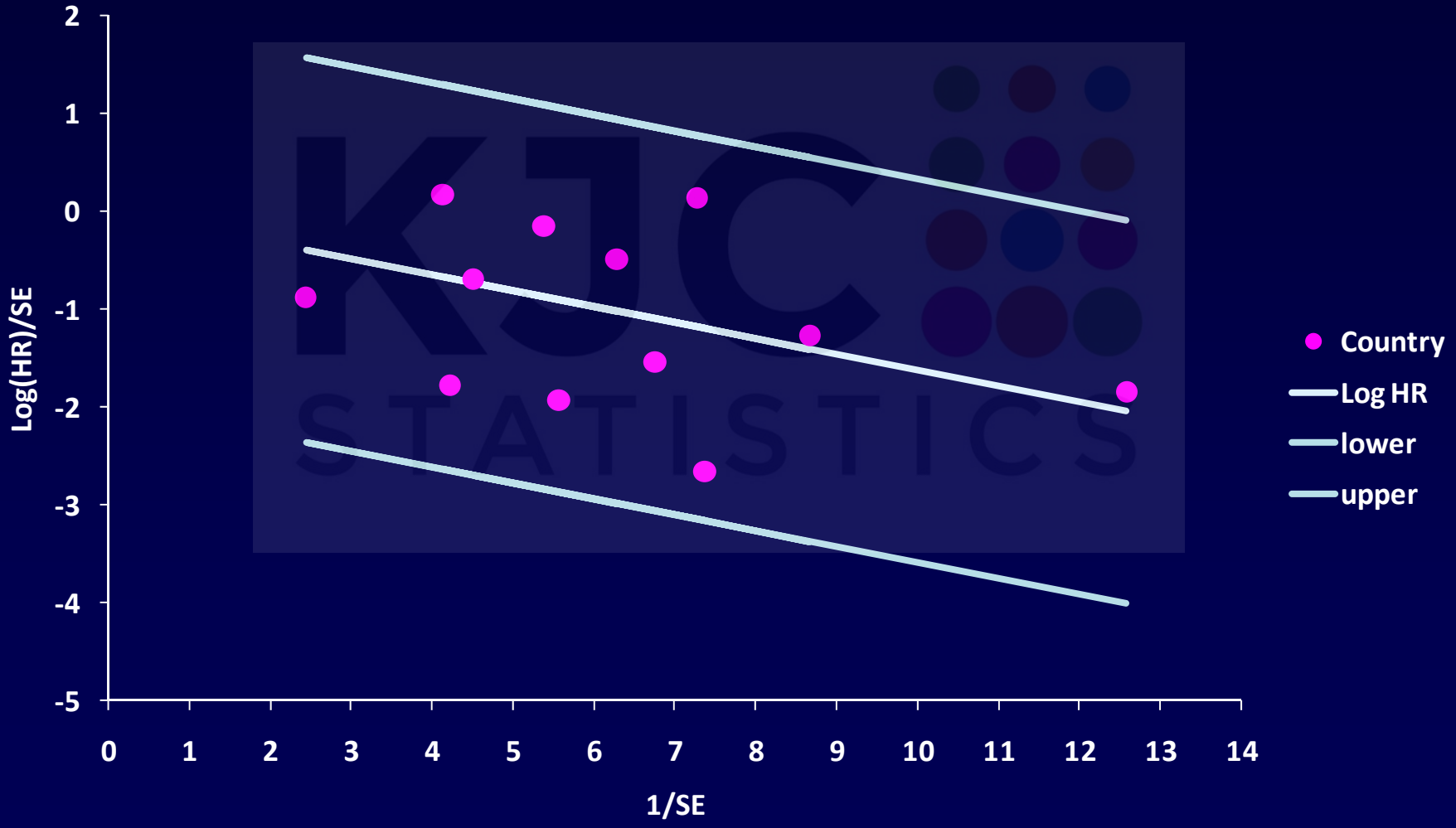
Given the overall result and based on distribution of events across countries, expect 2 countries to show effects > 0

Primary (mortality + hospitalisation)

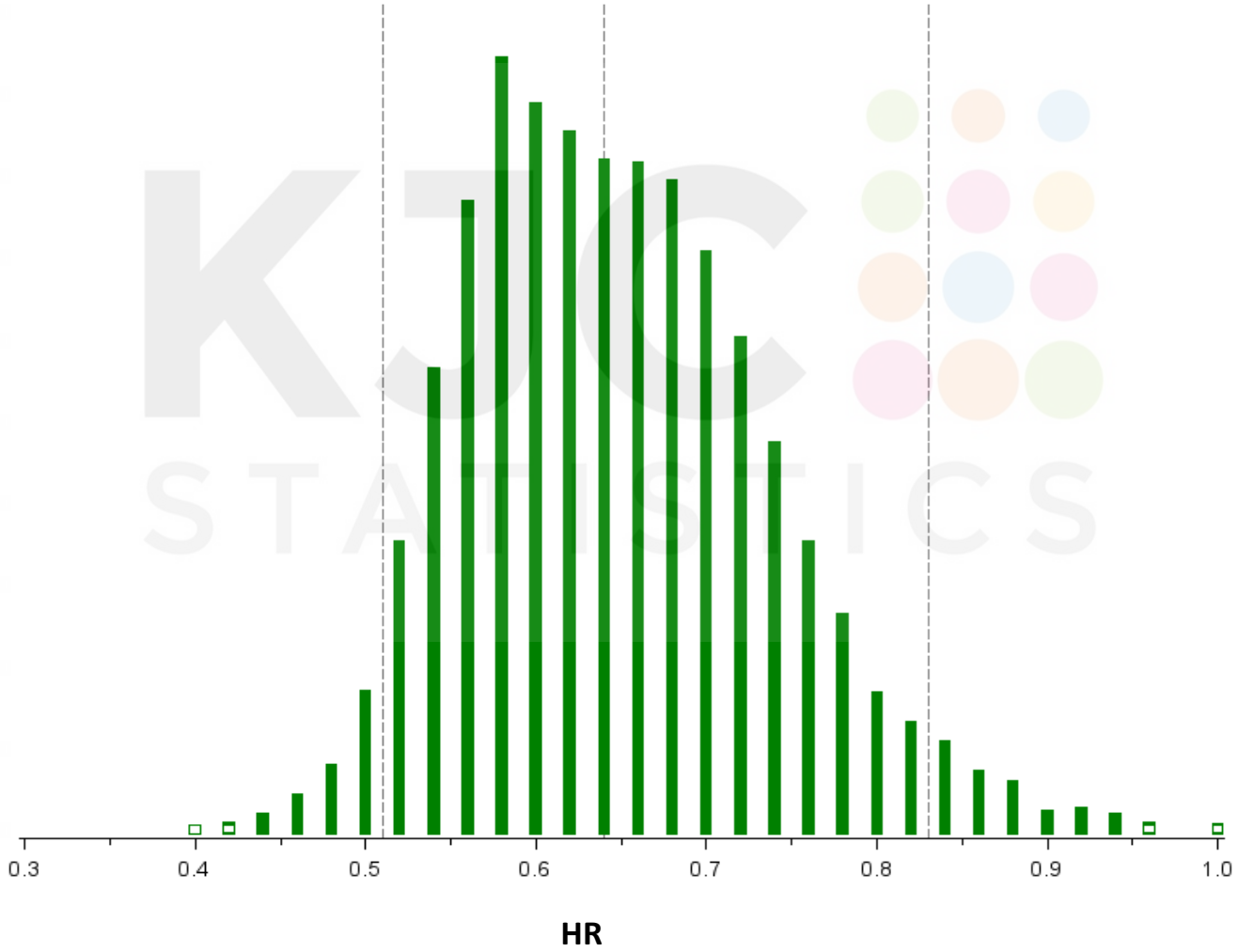
	Drug		placebo			
Country	Events	N	Events	N	vi	xi
belgium	31	68	31	66	0.0347	-0.0299
czech	35	123	50	124	0.0324	-0.3486
Denm/Fin	64	161	64	164	0.0189	0.0185
germany	88	252	100	247	0.0134	-0.1479
hungary	57	211	72	212	0.0220	-0.2289
iceland	6	19	10	22	0.1686	-0.3642
norway	41	97	48	105	0.0254	-0.0784
poland	26	102	25	102	0.0589	0.0392
sweden	15	39	27	46	0.0563	-0.4227
holl/swiz	68	299	95	291	0.0185	-0.3615
uk	26	87	29	83	0.0494	-0.1563
usa	184	532	216	539	0.0063	-0.1473

Given the overall result and based on distribution of events across countries, expect 2 countries to show effects > 0

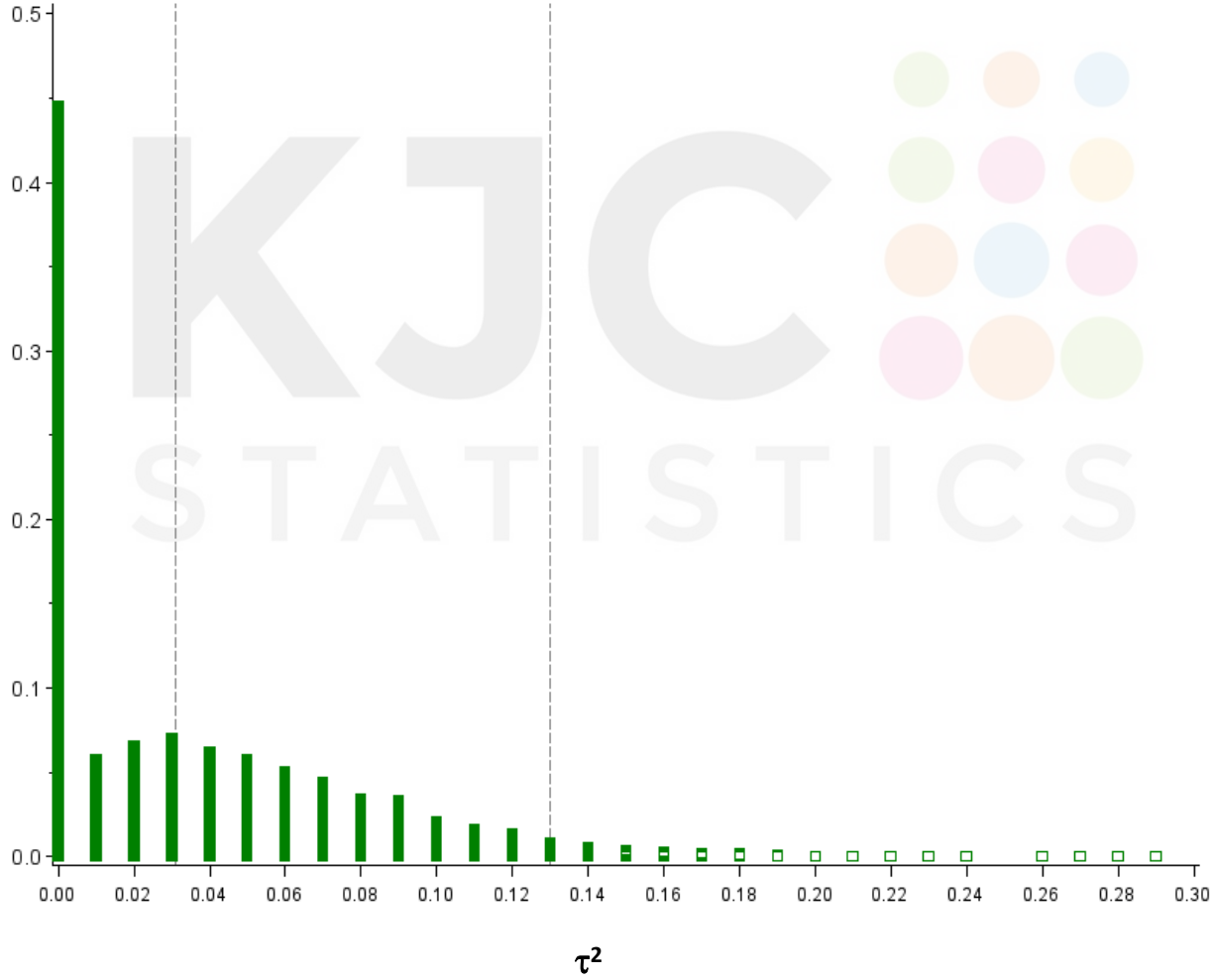
MERIT-HF: Primary (mortality + hospitalisaion): HR=0.85 95% CI (0.77. 0.93), τ^2=0, σ^2=0.024, I^2=0



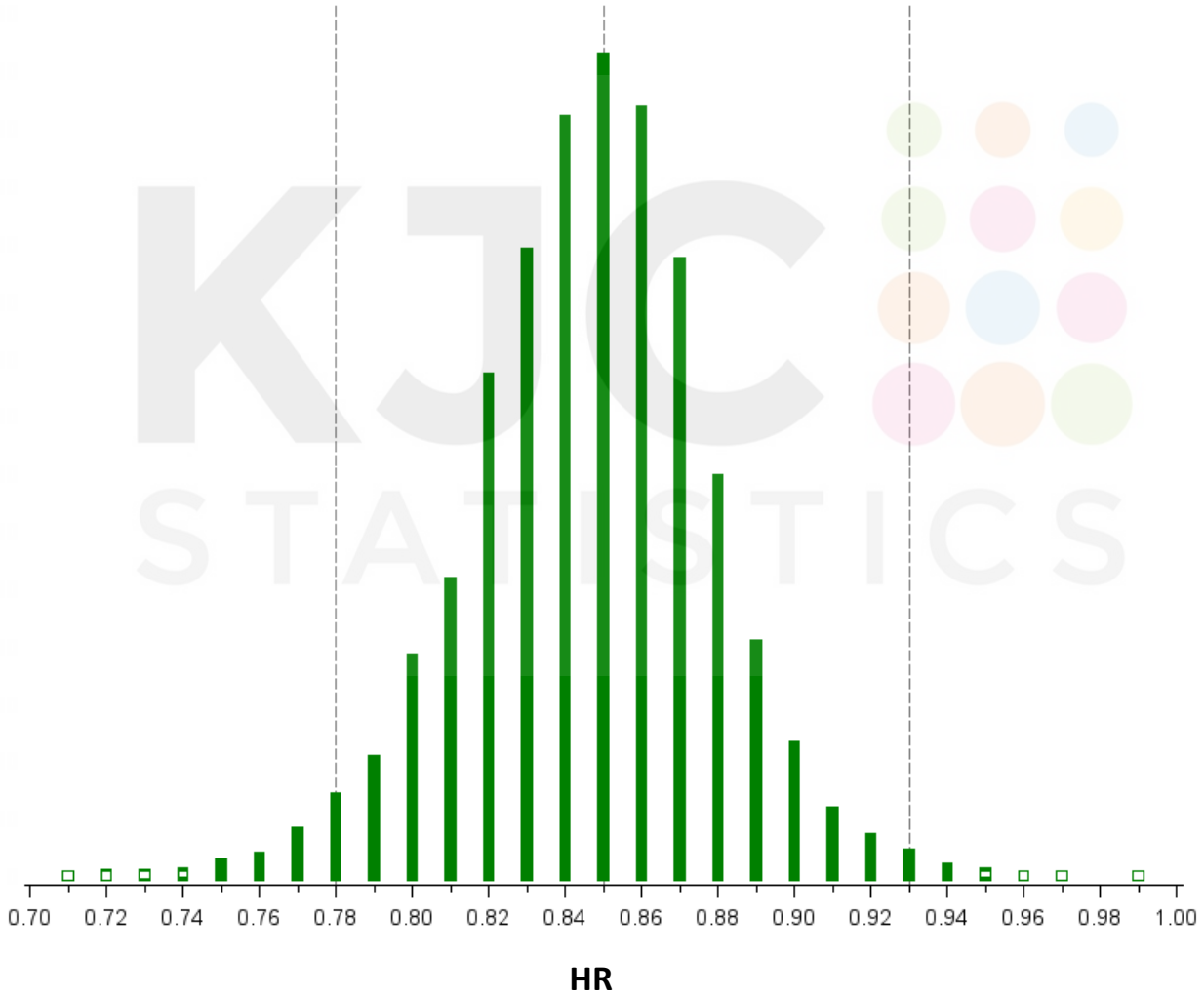
MERIT-HF:
Mortality: HR=0.64 95% CI (0.51. 0.83) by bootstrap



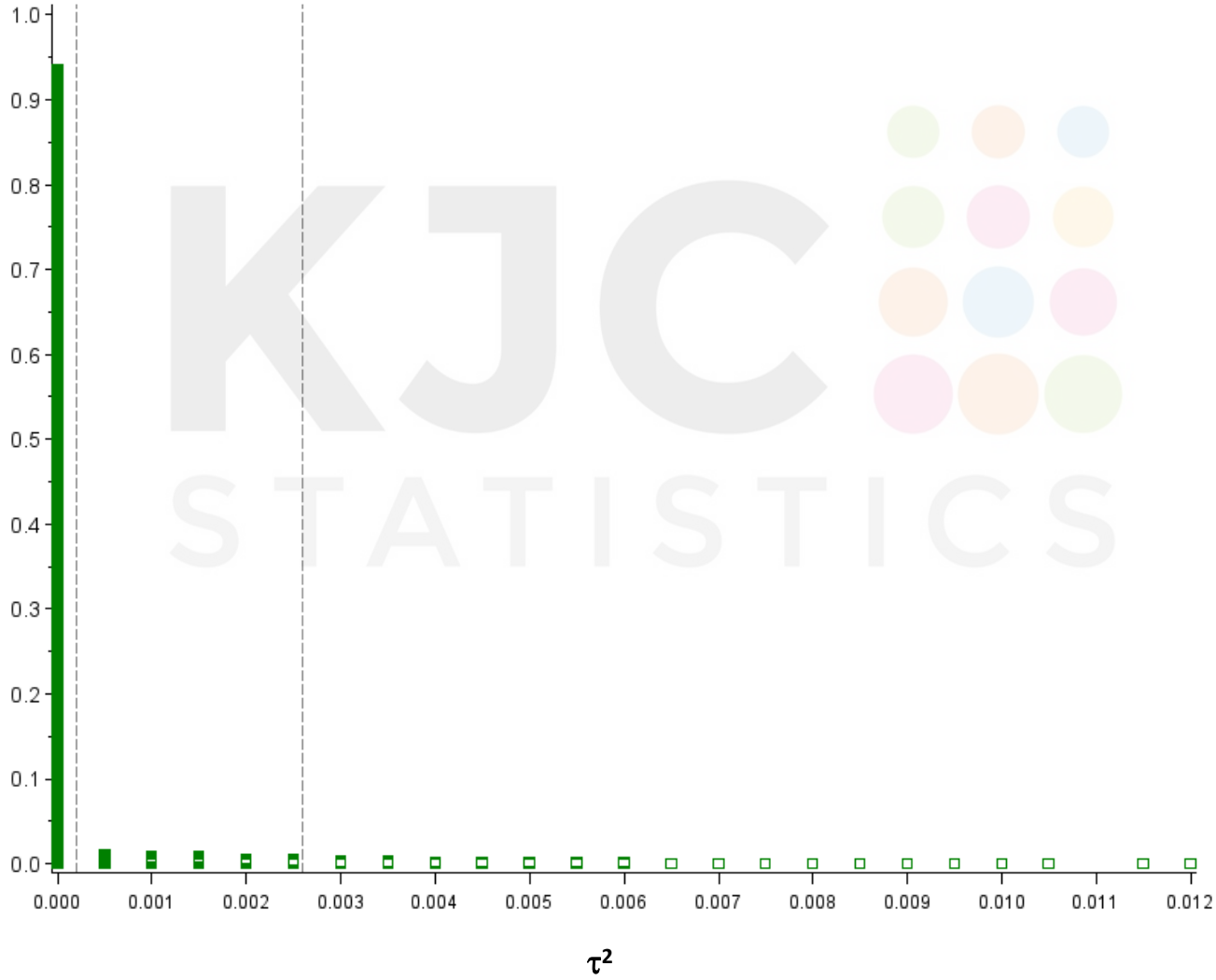
MERIT-HF:
Mortality: τ^2=0.031 95% CI (0. 0.13)



MERIT-HF: Primary (mortality + hospitalisation): HR=0.85 95% CI (0.79 to 0.90) by bootstrap



MERIT-HF: Primary (mortality + hospitalisaion): τ^2=0, 95% CI (0, 0.0025)



Wedel and DeMets, (Am Heart J 2001;142:502-11.)

- Just as we must be extremely cautious in over-interpreting positive effects in subgroups, even those that are predefined, we must also be cautious in focusing on subgroups with an apparent neutral or negative trend.
- We should examine subgroups to obtain a general sense of consistency, which is clearly the case in MERIT-HF.
- We should expect some variation of the treatment effect around the overall estimate as we examine a large number of subgroups because of small sample size in subgroups and chance.
- Thus the best estimate of the treatment effect on total mortality for any subgroup is the estimate of the hazard ratio for the overall trial.

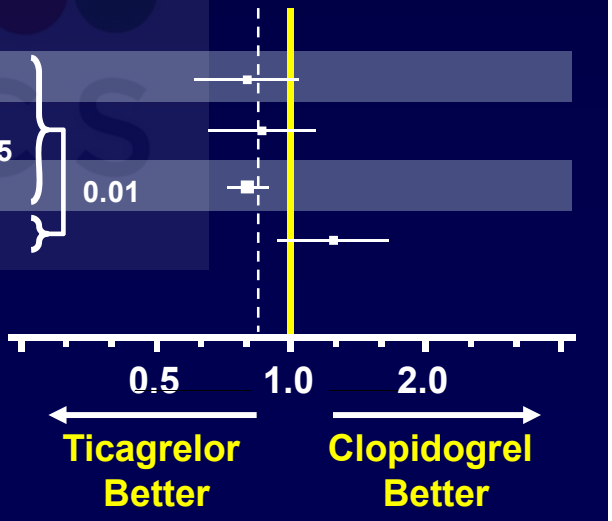
PLATO

- Randomized double-blind study comparing BRILINTA (N=9333) to clopidogrel (N=9291), both given in combination with aspirin, in patients with acute coronary syndromes.
- Primary endpoint was time to first occurrence of CV death, MI or stroke.
- Randomisation across 41 countries.
- Primary endpoint met for BRILINTA 9.8% vs 11.7% events HR = 0·84 95% CI 0·77–0·92]; p=0·0003.
- Benefit also seen in overall mortality 4.5% vs 5.9% events HR = 0·78 95% CI 0·69–0·89]; p=0·0003.

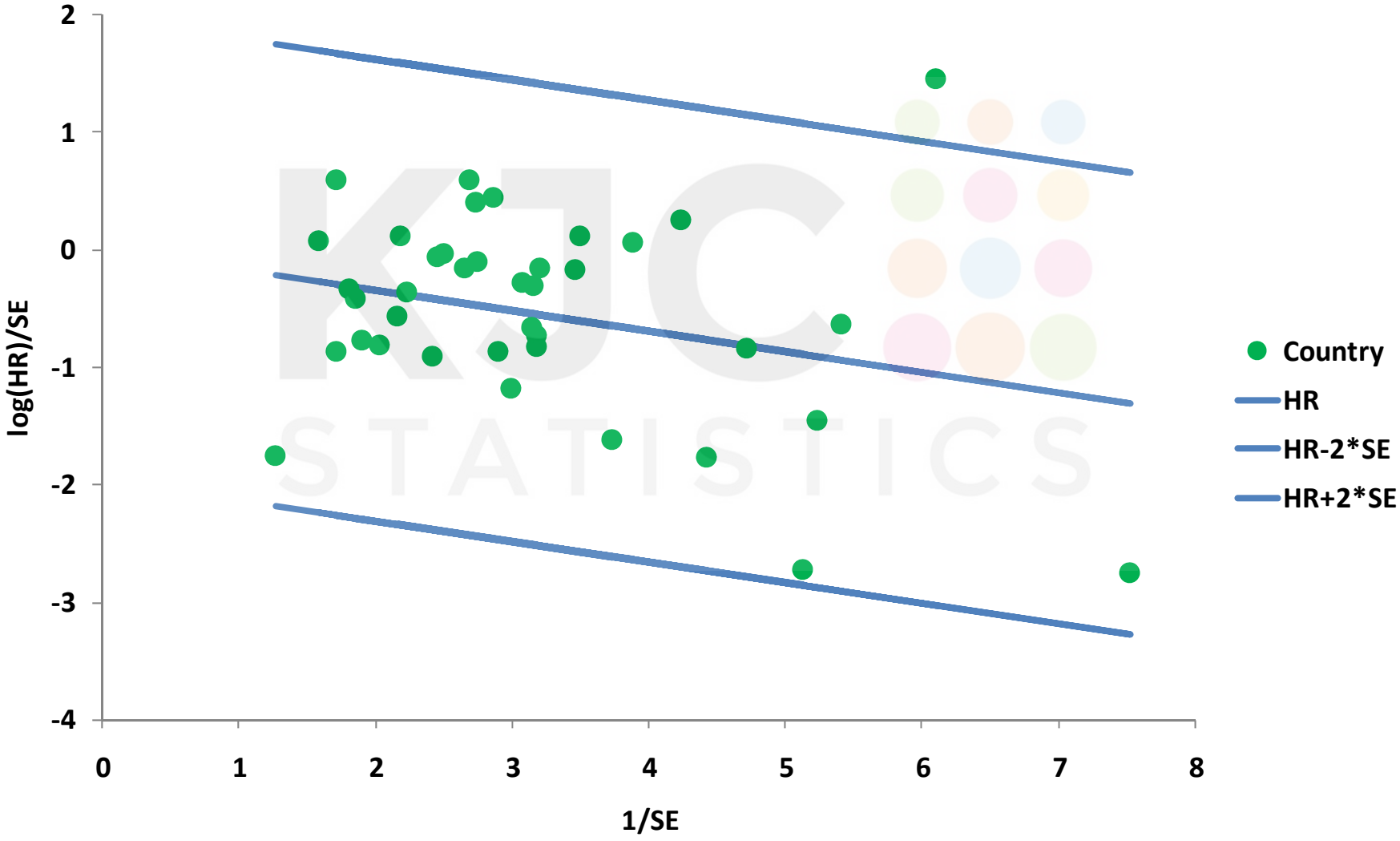
PLATO: Ticagrelor Effect Apparently Inconsistent Across Geographic Regions

- 31 pre-specified subgroup tests conducted for consistency
- No α-level adjustment for multiplicity
- Indication of qualitatively different outcomes by region
- Results in NA appear to be driven by US: HR 1.27 (0.92, 1.75)

Characteristic	Total Patients	KM at Month 12 Tic	KM at Month 12 Clop	HR (95% CI)	Interaction p-values
Geographic Region					
Asia / Australia	1714	11.4	14.8	0.80 (0.61, 1.04)	0.045
Cent / Sth America	1237	15.2	17.9	0.86 (0.65, 1.13)	
Euro / Md E / Afr	13859	8.8	11.0	0.80 (0.72, 0.90)	0.01
North America	1814	11.9	9.6	1.25 (0.93, 1.67)	



Galbraith plot



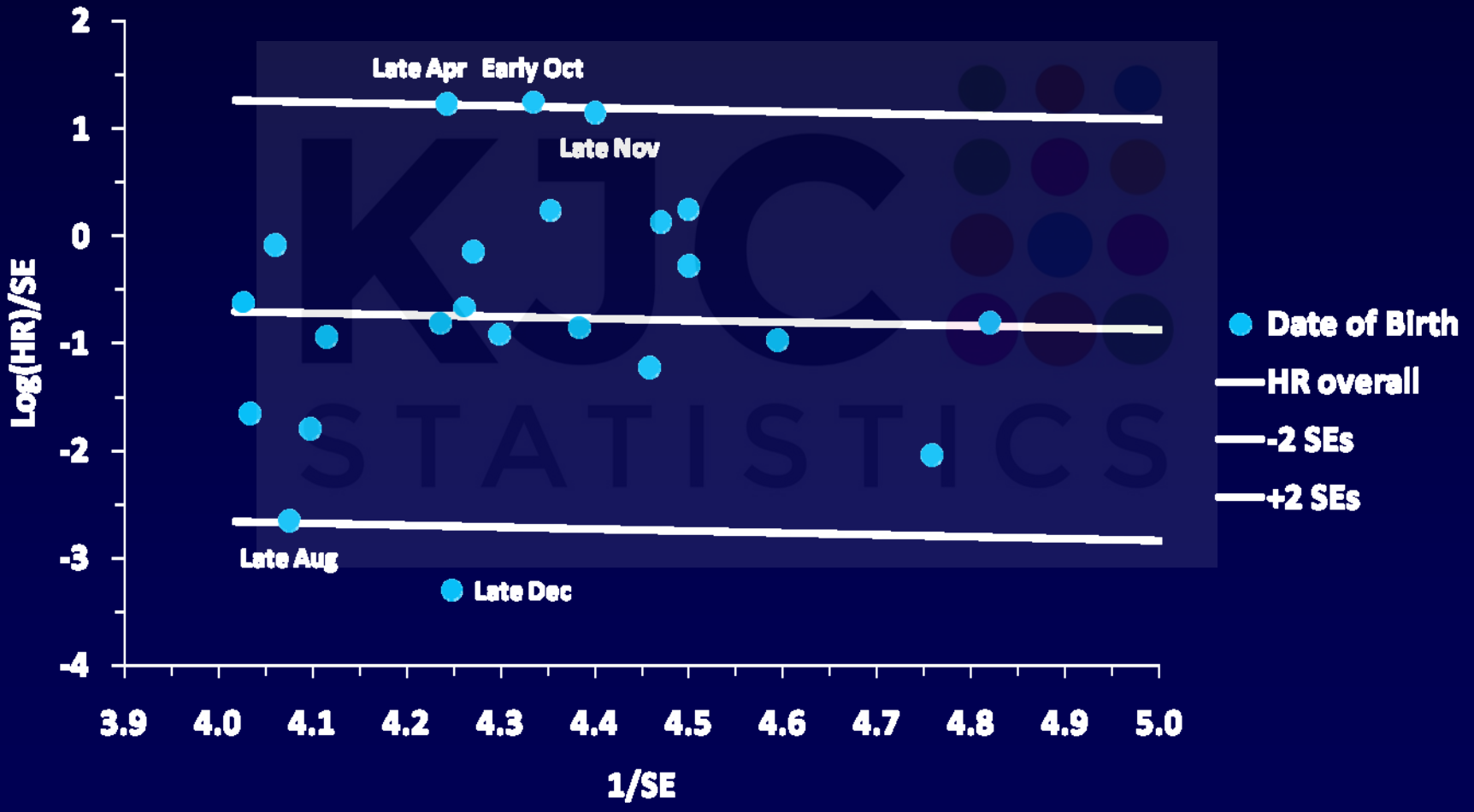
Probability of Observing At Least 1 Statistically Significant Treatment Interaction By Chance Alone is High

Correlation between tests	Fraction of simulations with at least 1 significant result in 31 tests	Equivalent adjusted p-value to retain overall false-positive error rate at 5%
0	79.1%	0.002
0.5	51.3%	0.004
0.9	17.3%	0.014
0.99	7.9%	0.031
1.0	5.0%	0.050

PLATO Pattern of effect reversals consistent with what would be expected in a large MRCT

Expected no. countries with HR >1	Actual no. countries with HR >1	Expected no. countries with HR >1.25	Actual no. countries with HR >1.25
12.9	12	6.2	3

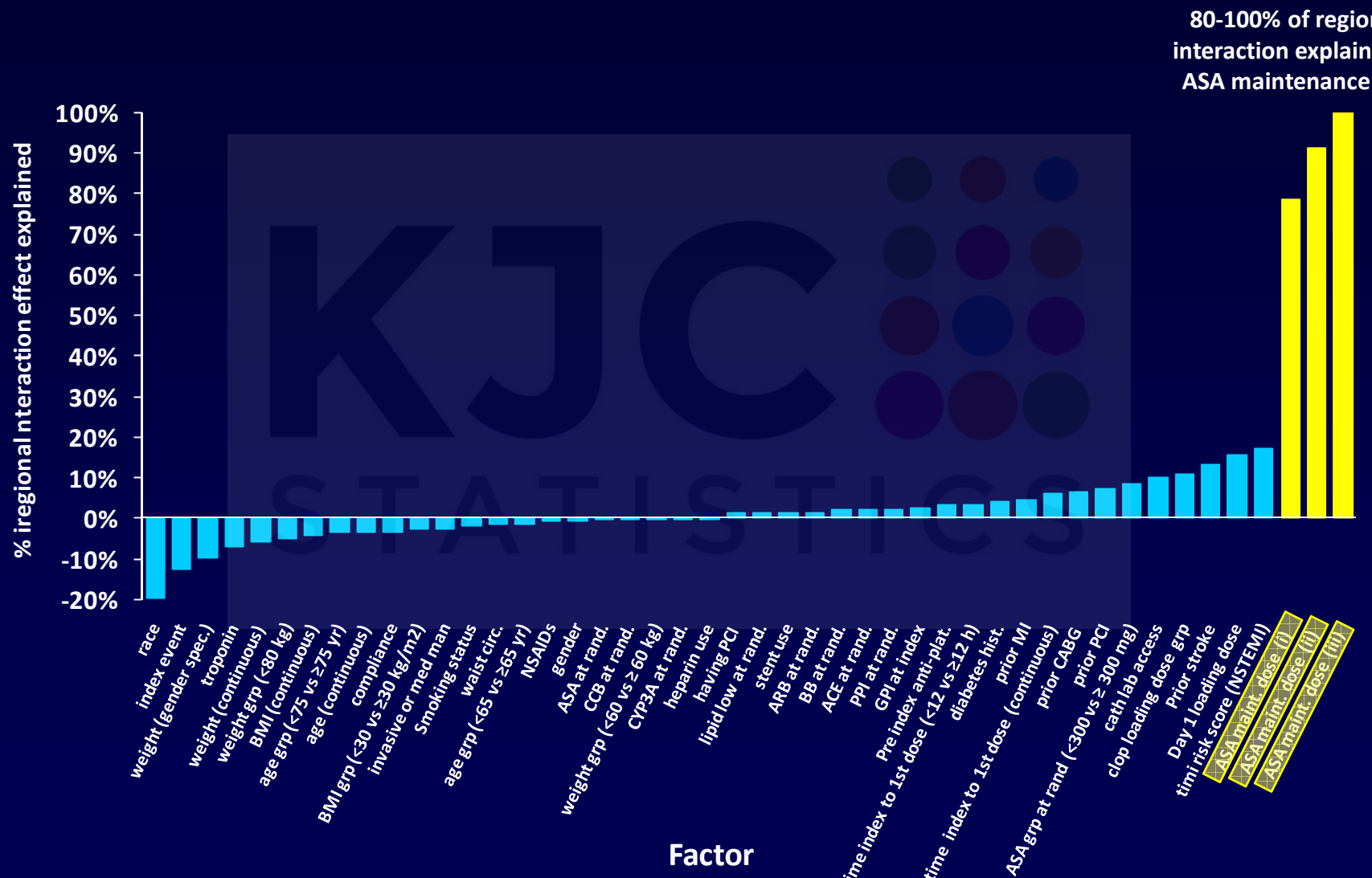
Analysis by date of birth



PLATO: What Kind of Factors or Patient Characteristics Might Explain the US vs Non-US Result?

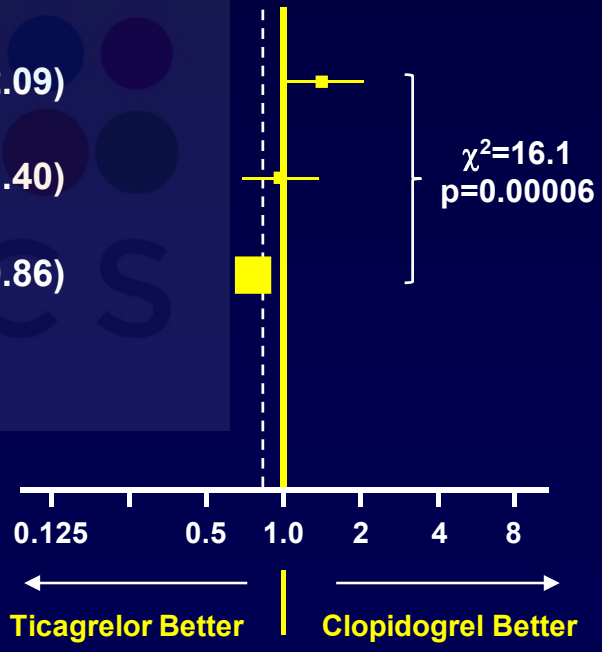
- **To explain a meaningful fraction of the US/non-US interaction, a factor is needed that simultaneously:**
  - **(i) has a strong qualitative interaction with randomized treatment for the primary endpoint and**
  - **(ii) is strongly imbalanced between US and non US settings**
- **Weakly imbalanced prognostic factors will likely not be sufficient to explain the US result**
- **Visual inspection for imbalances of clinical concern needs to be supported by an objective and statistically rigorous analysis of the data**

PLATO: No Factor Potentially Accounts for the Regional Interaction with the Exception of ASA Maintenance Dose During Therapy



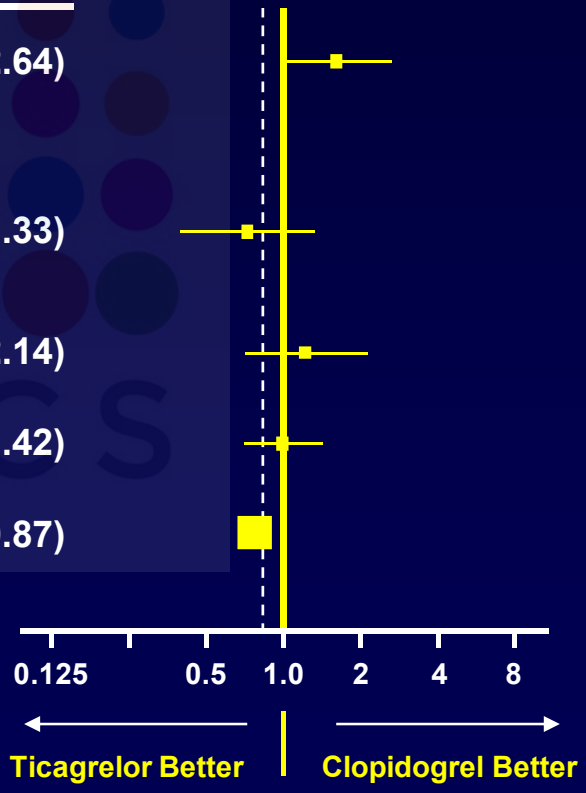
PLATO: The regional interaction is explained by an interaction with ASA maintenance dose

ASA Dose (mg)	Ticagrelor N	Ticagrelor E	Clopidogrel N	Clopidogrel E	HR (95% CI)
Overall					
≥300	464	68	492	50	1.45 (1.01, 2.09)
>100 – <300	525	64	527	65	0.99 (0.70, 1.40)
≤100	7733	565	7706	723	0.77 (0.69, 0.86)



PLATO: Similar Pattern of Treatment Effects in Relation to ASA Maintenance Dose in US and Non-US

Region	ASA Dose (mg)	Ticagrelor N	Ticagrelor E	Clopidogrel N	Clopidogrel E	HR (95% CI)
US	≥300	324	40	352	27	1.62 (0.99, 2.64)
	>100 – <300	22	2	16	2	
	≤100	284	19	263	24	0.73 (0.40, 1.33)
Non-US	≥300	140	28	140	23	1.23 (0.71, 2.14)
	>100 – <300	503	62	511	63	1.00 (0.71, 1.42)
	≤100	7449	546	7443	699	0.78 (0.69, 0.87)



Summary (1)

- MRCTs are essential in modern day drug development
- Some degree of observed variability is inevitable between regions – in a trial with 80% power and 4 equally sized regions, the chance of at least one reversal is ~30% and with 6 regions the chance is >50%. The corresponding figures are ~20% and 7 regions for a trial with 90% power.
- By definition, MRCTs are not designed, powered or intended to look statistically for true differences between regional effects – all the caveats, biases and pitfalls of subgroup analyses apply.
- Assuming at the planning stage that there is true heterogeneity in the treatment effect by region so ($\tau^2 > 0$) is problematic
  - What value to choose for τ^2 and on what basis? At what point does the MRCT become meaningless and impossible to interpret clinically due to excessive variability?
  - Sample size inflation will quickly render the trial infeasible.

Summary (2)

- A random effects analysis is not the answer especially if applying a Follmann adjustment
- More thought needs to be given to the consistency in defining 'region' across trials – "Brazil, Chile, Argentina and New Zealand"
- The optimum allocation of N across regions / countries and the criteria for 'consistency' are two sides of the same coin.
- For multiple regions, simplest, but least realistic, criteria is to require all regional effects > 0.
- For reference vs non-reference regions, most pragmatic criteria is to require 3 regions, is to require the point estimate for the reference region to be no less than the lower 100(1-2α)% CI estimate for the treatment effect in the total trial population.
  - provides some reassurance the treatment effect in the reference region is not worryingly less than the treatment effect seen in the total trial population.

Summary (3)

- Graphical methods are very informative when examining regional effects – Galbraith, Q-Q and forts plots are essential.
- Decomposition of the overall χ^2 into a weighted sum of t^2 statistics for i^{th} region vs all others is very informative and should be done routinely.
- If there is truly believed to be a region with a different result, extensive evaluation is required to assess play of chance, and to look for possible confounders in medical practice, medical care, quality of trial conduct, cultural, biologic and genetic factors that may be explanatory.
- In the end "Thus the best estimate of the treatment effect ... for any subgroup is the estimate of the ... [treatment effect] for the overall trial" Wedel and DeMets