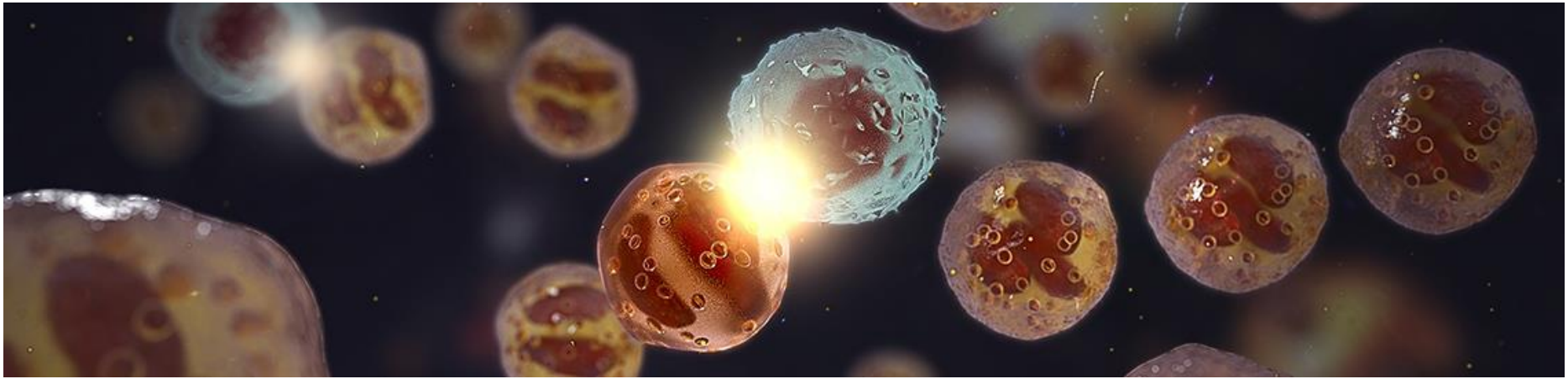
Consistency of treatment effect across pre-specified subgroups – should we (and, if so, how?) adjust for biases?

Speaker: David Svensson – Advanced Analytics, AstraZeneca R&D Sweden.

PSI Amsterdam June 2018

18 June 05



DISCLAIMER

- The **opinions** expressed in this presentation are those of the **authors,** and do **not** necessarily reflect the official policy of AstraZeneca.

Content & Outline

- A key step in the evaluation of any pivotal Ph III RCT is to make a risk-benefit assessment and **identify the right patient population** to treat.
- This is '**consistency of treatment effect**', a regulatory requirement (e.g., ICHE5, E9 & E17, EMA [13, 21]). A failure might lead to restricted labelling.
- Special case of interest: regions/countries in MRCTs. Overall result versus country-specific results?
- Despite the importance, no detailed regulatory guidance how to statistically demonstrate consistency. Current standard practices: **statistical issues**.
- **Alternative methods** have been proposed: some discussed here.
  - (*industry view on these*?)

Short remarks:

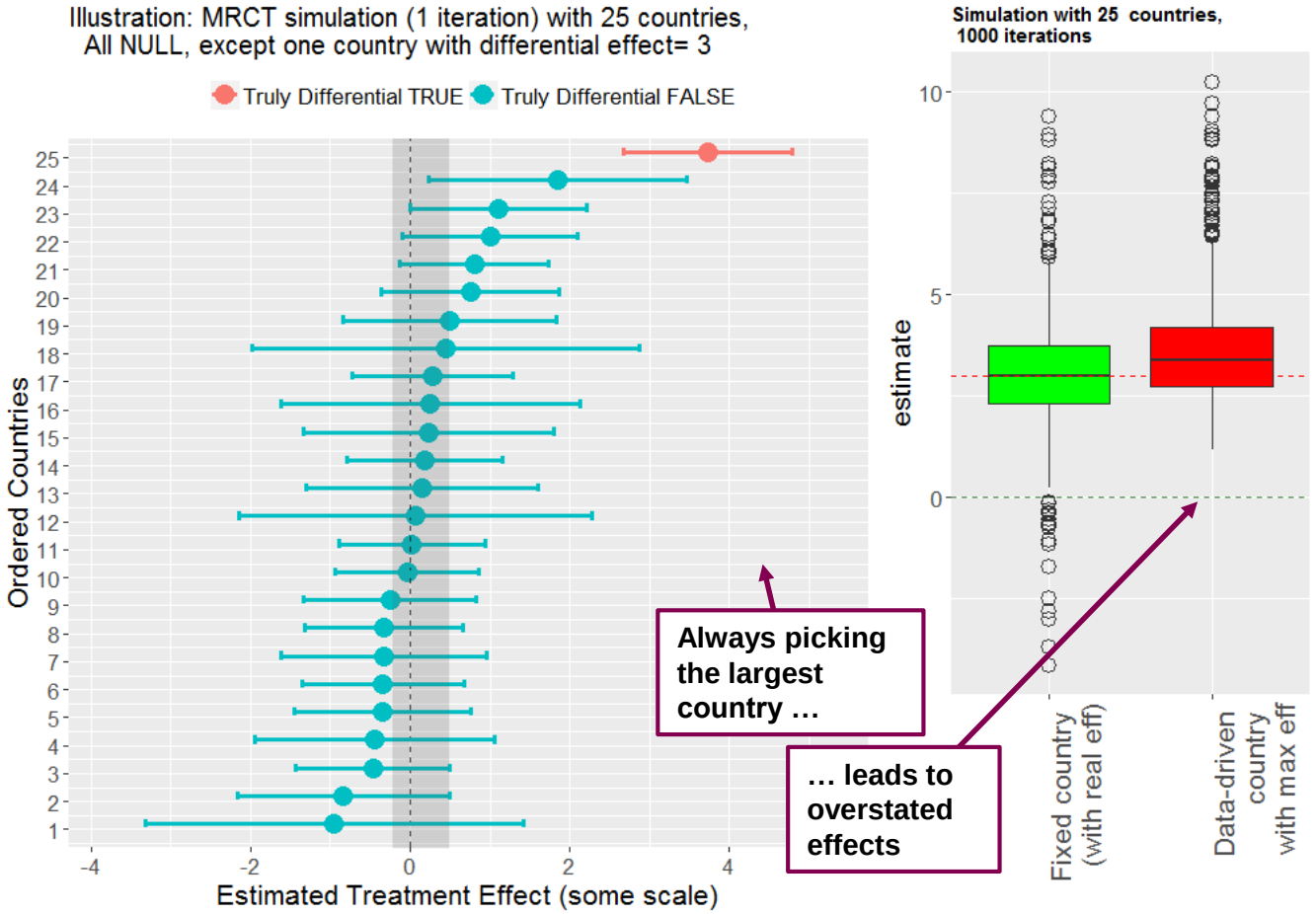
- We have (at this time) **no** strong preference regarding proposed approaches.
  - Rather: are there **any good alternatives** out there?
- But focusing mostly on **shrinkage** here (as an example of a new approach).
- (Aspects will be illustrated with simulations: **exaggarated** effects for clarity).
- Note**: Overlapping** versus **non-overlapping** subgroups:
  - Mostly non-overlapping here (= *a patient belongs to one subgroup only*).
  - E.g., ”country” in MRCT.
- The overlapping (general) case is harder for various reasons (more later).

Are the standard estimates biased?

- Not per se ..
- … unless the reason you look at it is data-driven.

Any data-driven selection: bias.
(e.g., [1])

True even if a country truly has larger expected value (simulation illustration).



Current Approach to consistency: issues …

We reckon that consistency asessement is not merely a statistical exercise, (other considerations too, complex …)

Still, **expect more attention to extremes**; i.e., biases are inherent.

Subgroup estimates = limited data = **large variability**. (Risk of false positives).

Testing sometimes done: e.g., Global Interaction Test (GIT)?

- Low power (trials not sized for it).
- Multiplicity (If a test per subgroup factor – general subgroups).
- Not a testing problem. *Absence of evidence* vs *evidence of absence* …

Well-known issues [2], [5], [29], and alternatives have been proposed …

Any more holistic approach available?

- E.g., **EMA guideline** [13]: useful to assess subgroups together (display full range) although not many details given. Recommendation:

  1. Aim to **pre-group subgroups** into categories (scientific credibility reasons),
     - (A) differential effect plausible, or (B) no differential effect expected.

  2. Use **graphical** methods (e.g., Forest plots).

- **EFSPI Subgroup WG** – White Paper [6]; compared various methods (e.g., simulations); e.g., SEAMOS [8], SIDES [24], Bootstrap Bias Reduction [33], GIT. Recommendation: as EMA, plus:

  – Consider **expected-worst under NULL** [8], to provide a context of the results.

- Jury still out on some other methods? (E.g., Bayesian shrinkage?)

Bayesian Shrinkage has been suggested, e.g.,:

- *Empirical shrinkage estimator for consistency assessment of treatment effects in multi-regional clinical trials.* [30].
- *Multi-regional clinical trial design and consistency assessment of treatment effects.* [31].
- *Treatment effect heterogeneity for univariate subgroup in clinical trials: Shrinkage, standardization, or else.* [35]
- *Exploratory Subgroup Analyses in Clinical Trials.* [32]
- *Bayesian models for subgroup analysis in clinical trials.* [23]
- *Comparing Approaches to Treatment Effects Estimation for Subgroups In Clinical Trials*. [26].
- *Bayesian Assessment ot the Influence and Interaction Conditions in Multipopulation Tailoring Clinical Trials*. [27]
- *A Bayesian Approach to Evaluating Regional Treatment Effect in a Multiregional Trial*. [3]

The idea of Shrinkage

- Well-known **classical** concept, e.g., [11, 12, 26, 36].
  - Stein's classical 'shocking' theorem:
  - *ML estimates can be (dramatically) improved (sometimes)*. [11]

- **Widely used** for high-dim data, e.g.,
  - Microarray screening [11], Pharmacovigilance FDA Signal Detection [19].

- **The Core Idea**: seems to be:
  - *True effects less spread out than observed ones…*
  - *Some shared information across subgroups.*
  - *Self-tuning smoothing towards overall estimate...*

- Borrowing information might help with sparse subgroups (e.g., few events)?

Shrinkage, basic model

i = countries: standard est. $\hat{\delta}_i \sim N(\delta_i, s_i^2)$

Prior for (true) country effects:

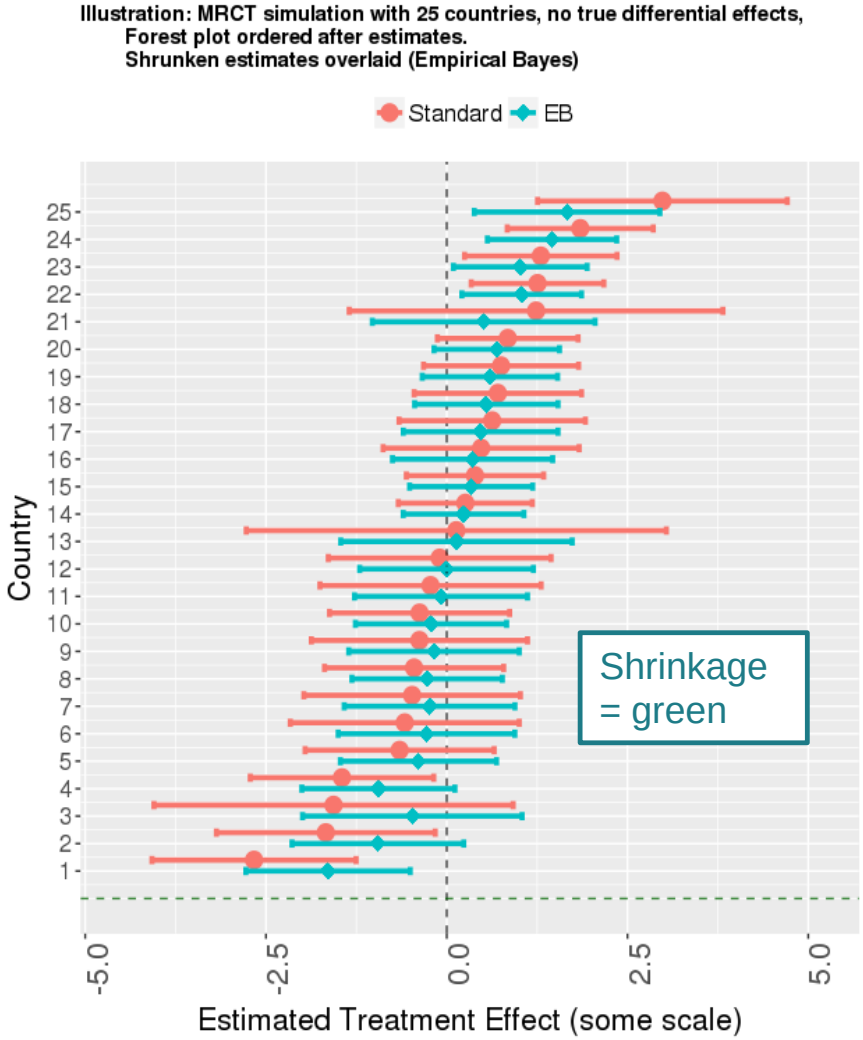
$$\delta_i \sim N(\delta, \tau^2)$$

Gives posterior 'estimate' (MAP):

$$\tilde{\delta}_i = w_i \cdot \hat{\delta}_i + (1 - w_i) \cdot \tilde{\delta}$$

with $\tilde{\delta}$=overall (RE) estimate, and weights: $w = \tau^2/(\tau^2 + s_i^2)$

Model fit? MCMC (Full Bayesian [16]), or **Empirical Bayes** (ML fit of prior to data [26, 35]).



Drivers of amount of Shrinkage? (How much modified):

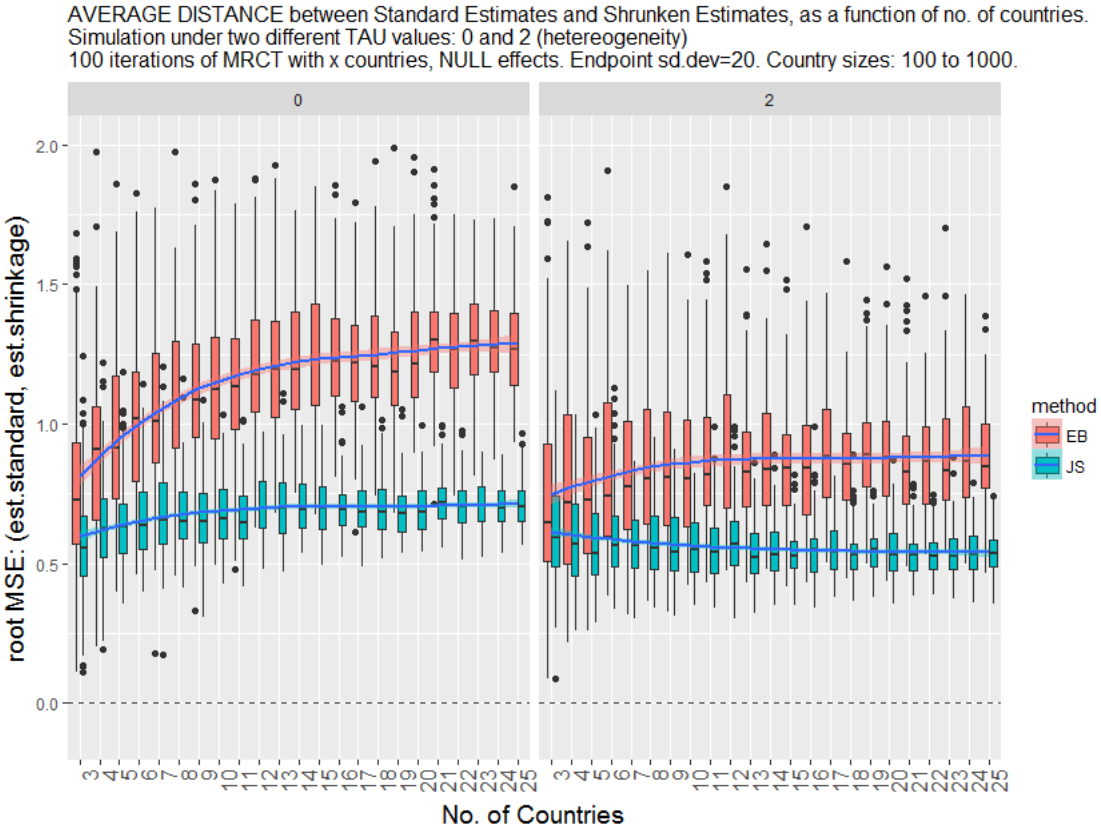
- **Between $\hat{\tau}$ vs within s_i^2**

$$s_i^2 \approx 0 \;\Rightarrow\; \tilde{\delta}_i \approx \hat{\delta}_i$$

$$s_i^2 \; large \;\Rightarrow\; \tilde{\delta}_i \approx \tilde{\delta}$$

- **No. of countries**
  (=x-axis in graph)

- Shrinkage **method** choice
  (several exist –more later).



Some remarks on Emp Bayes estimates

- Remark 1: They are **unbiased** (despite shrinkage towards overall – any confusion?). [under the correct model]

- Remark 2: They also do **bias-adjustments** of 'random-highs':
  – The max.EB (and min) is still biased, but less so.

- Remark 3: Modified ('Improved' or 'manipulated'?) estimates: **controversial**?
  – Assumes 'exchangability' ($\delta_i \sim N(\delta, \tau^2)$: unrealistic?).
  – (Note: an assumption re. the *unobservable*, *true* country effects).

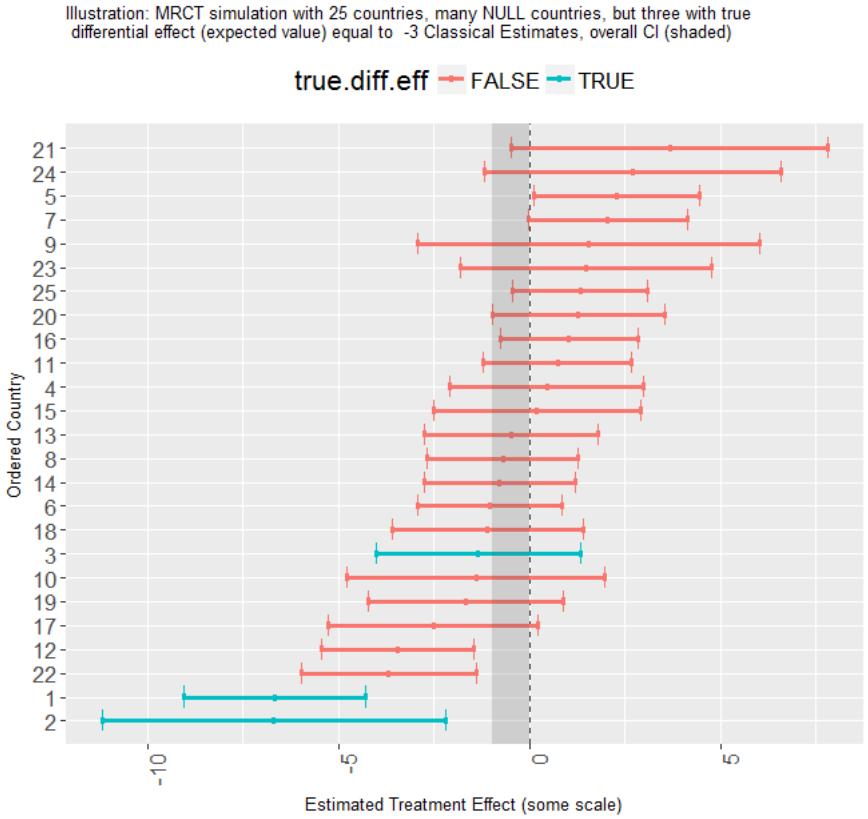
Real differential effects will be shrunken - but recall, noise shrunken too.
  – Does it make it ***easier* or *harder*** to detect truly differential subgroups?

Question: *shrinkage useful as a kind of secondary 'sensitivity analysis*'?

Simulation illustration, under broken assumtion:

What if exchangability is violated?!

- Iterations of simulation of
  - MRCT with 25 countries.
  - 22 with **NULL** effects, δ_i=0.
  - 3 **DIFF** countries with δ_i=x. (-3 in graph)
- **QUESTION**: our ability to detect this, with/without shrinkage?
- Tracking MSE & BIAS & Interval Coverage.
- (Also overlaps: country-specific intervals relative to the overall interval).



Some results, broken assumption (highlights only):

(Full results in Back-Up section)

Considerably **lower MSE for** Emp.Bayes, overall seen.

- (but not uniformly so)..

EmpBay **estimates biased**:

- differential countries were under-estimated.
- The null countries (zero effect) were slightly over-estimated.

But what does it mean, in terms of ability to detect differential countries?

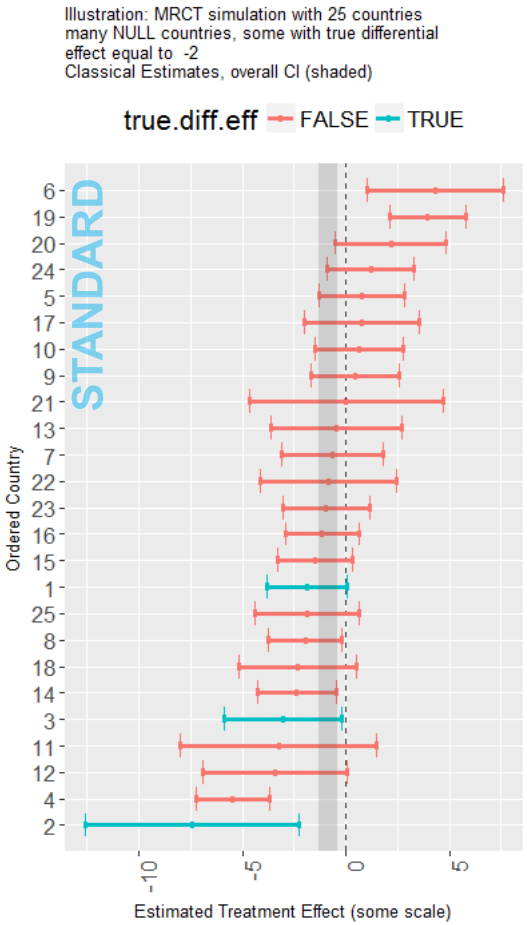
CI overlaps? (under broken assumption).

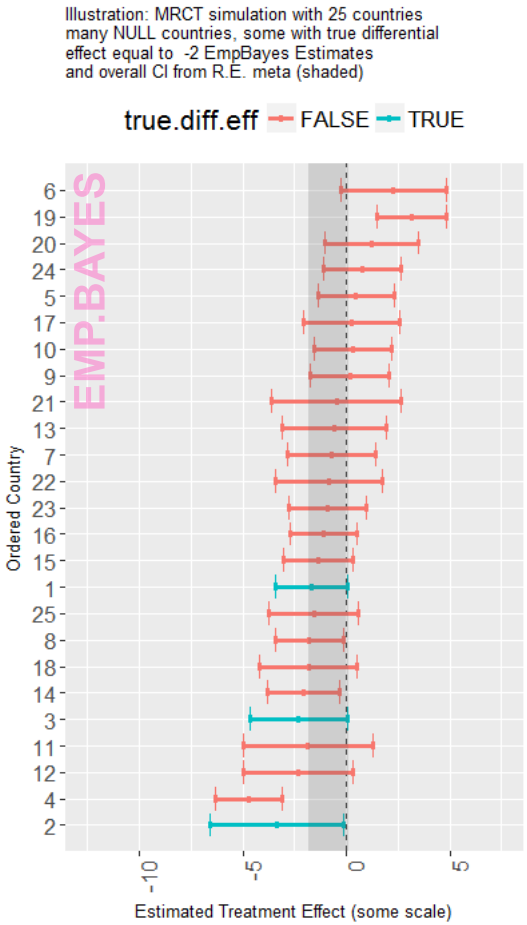
What if **CI overlap** was the key aspect? (Discovery Rates).

- **QUESTION**: ability to detect differential countries via 'non-overlaps', BEFORE/AFTER shrinkage?

1. $\hat{\delta}_j^{EB} \rightarrow$ center (estimates moving).
2. $\|CI_{ac}^{EB}\| \geq \|CI_{ac}\|$ (width, allcomer).
3. $\|CI_j^{EB}\| \leq \|CI_j\|$ (width, countries j).

How will it play out? (1,2 vs 3)



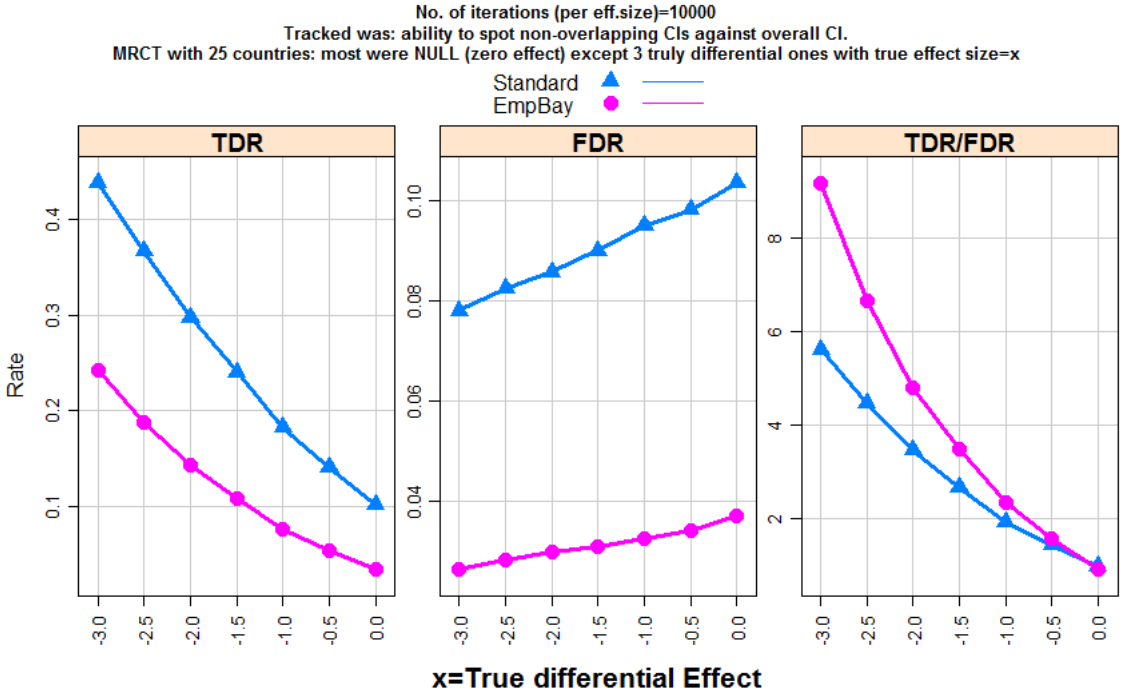


Easier or harder: preliminary results:

Many iterations,
each set with true
differential effect δ_i= x.

Tracked (for both methods):

- **T**rue **D**iscovery **R**ate:
  - (true countries detected?)
- **F**alse **D**iscovery **R**ate:
  - (NULL countries detected?)



Conclusion:
Ratio better ”after shrinkage”, but driven by lower FDR.

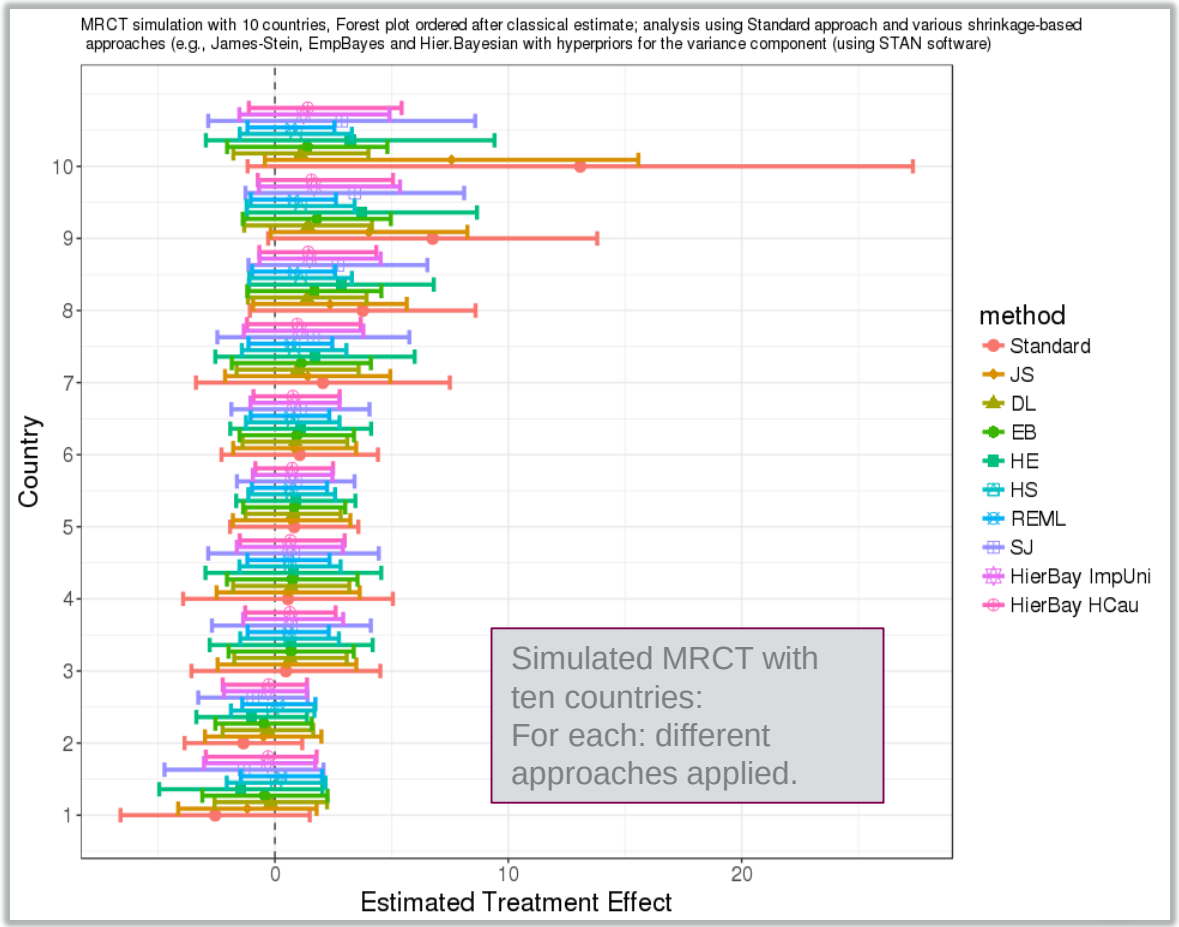
Shrinkage: not just Emp.Bayes; there is more …

- Empirical Bayes is only one of the possible forms of Shrinkage.
- **Meta Analysis** approaches (same model, but different 'τ^2 approaches' [28,37])
- Other well-known instances: **James-Stein** (frequentist approach) [30].
- **Full Bayesian Hierarchical** Model [16]:
  - Don't fit the prior to the data (as EB did): instead, 'let the data speak'.
  - But requires hyperpriors for variance component
  - Computational intensive fitting (e.g., STAN, hamiltonian chains [18]).
- (One motivation: EB ignores uncertainty in prior estimation, FB doesn't).

Shrinkage: many flavours, different results...

(RE approach, various options [37, 27], Hier. Bayes [16,17], Emp.Bayes [37], James-Stein [30]).

- Which is 'Vanilla'?
- Sponsor cherry-picking?
- Note:
- Amount of shrinkage.
- Interval width.

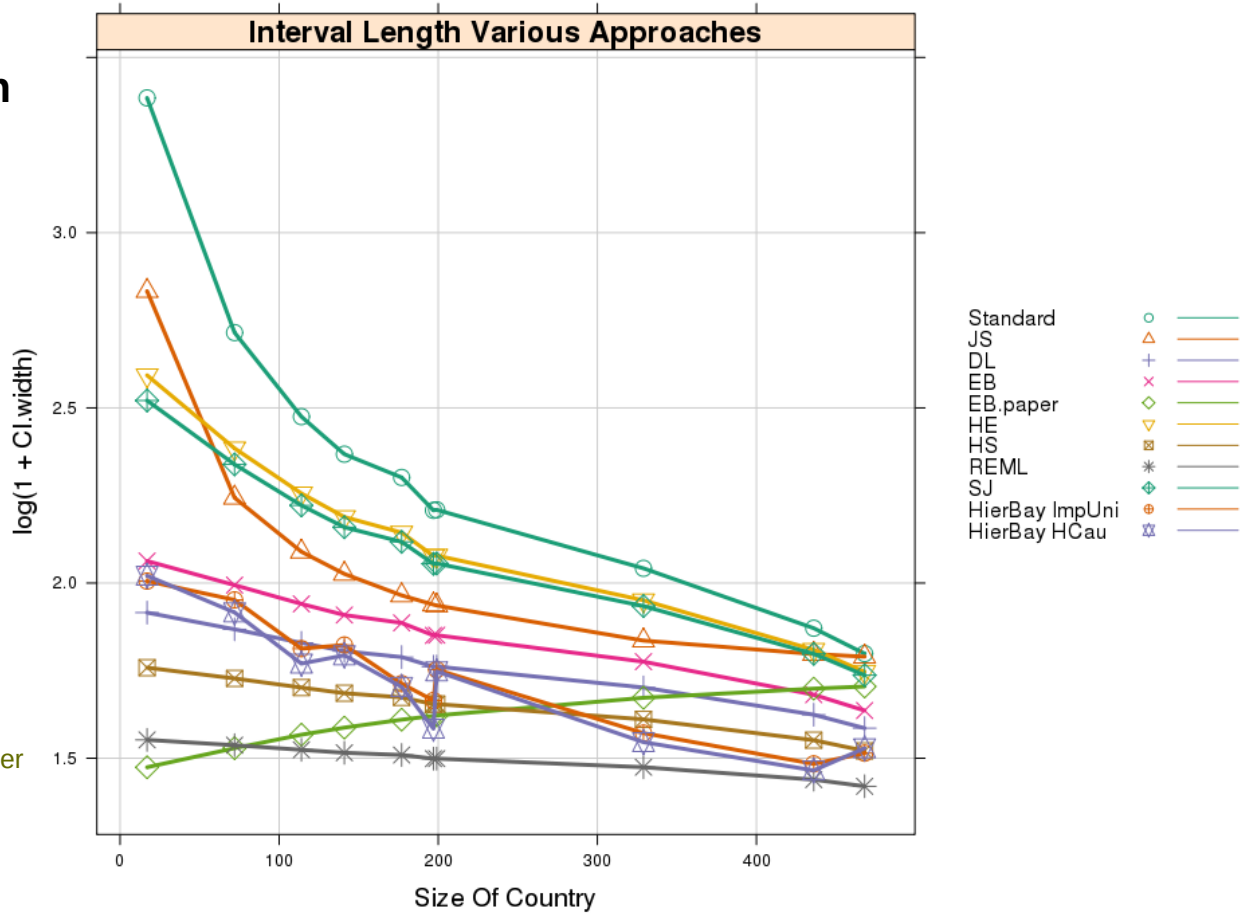


Shrinkage flavours; operating char. differ …

- Interval Length as a **function of subgroup size** – across different shrinkage flavours.

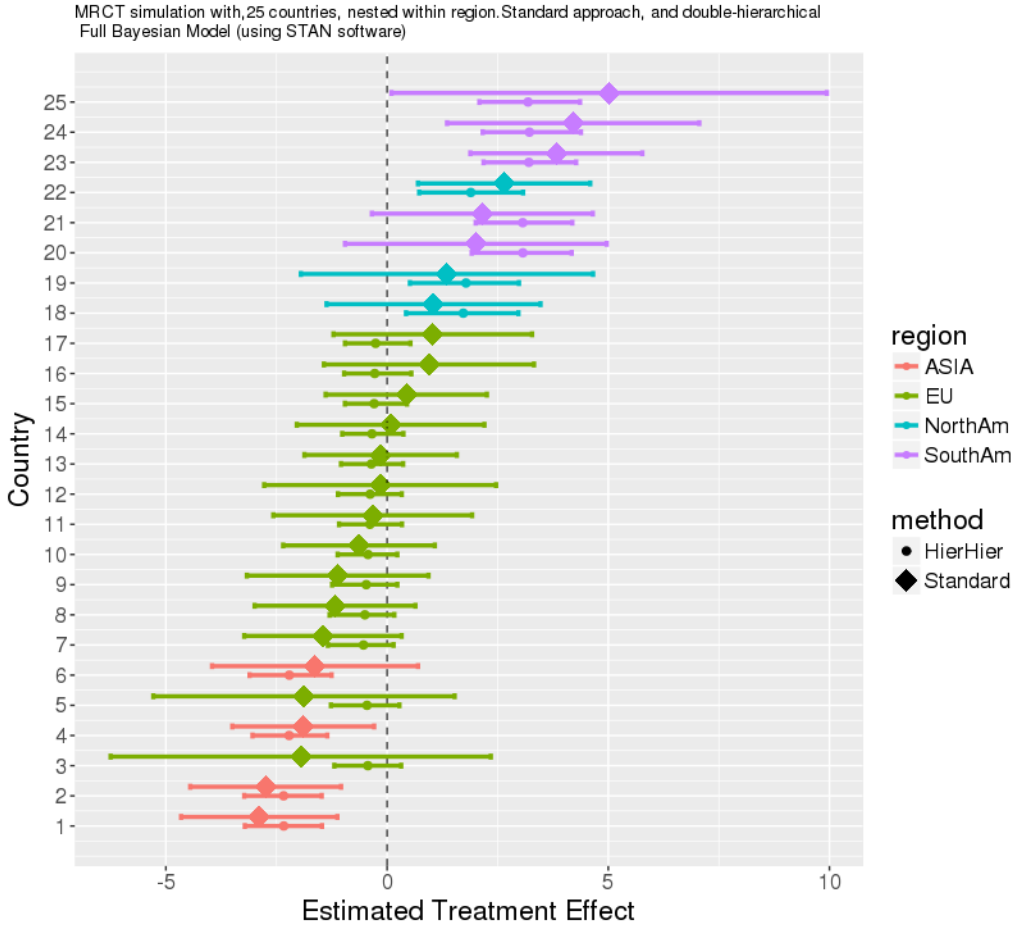
(RE approach, various options [37, 27], Hier. Bayes [16,17], Emp.Bayes [37, 30], James-Stein [30]).

(EmpBayes Variance formula wrong in [30]: smaller regions got more precise CIs …)



Multi-level hierarchical (e.g., country within region?).

- In the spirit of the EMA guideline recommendation [13, 7]; **shrinkage within pre-specified stratas** useful?
  - Could be 'region'.
  - Or, other cathegories:
    - Cat1: differential eff. not expected
    - Cat2: differential eff. plausible
- Possible using a Full Bayesian model.
- Assumptions:
  - Relaxed assumption re. country exchangability.
  - Prior for Regional eff.,
  - Prior for country-eff. within Regions,
  - Hyper-priors for variance components.
  - Unknown operating characteristics?



Shrinkage; overlapping subgroups:

The general (overlapping) case (e.g., `Disease Status, Conc.Med, Gender`):

- Hierarchical Bayesian models after splitting patients into **mututally exclusive** (disjoint) subgroups (e.g,. [9, 26, 38]).
  - Sometimes model fitting instability due to 'sparse cells'?
- Special case of Model selection: **Model averaging** of subgroup-specific models; [1]. Primarily developed for bias-reduction of *selected* subgroups.
  - But does shrinkage for all subgroups (via BIC model uncertainty).
  - Technically different: ensemble of models instead of a single model (such as e.g., with the EB)

Shrinkage; some question marks:

- Shrinkage **assumes what we want to demonstrate?**
-
  - e.g., FDA [34] – ”*apriori assuming exchangability”.*
  - *(”Perhaps this can be helpful in limiting […] random highs*”).

(Stratified assumption/mixture more realistic?).

- **Normality** assumption of the true effects:? Some **skewness** and/or heavier tails more realistic?

- Full Bayes **requires hyper-prior**:
  - No non-informative exists (variance component). [17]. So, which is it?
  - Diagnostics – **operator dependent**?

Conclusions:

- The current approach has **flaws**, and novel approaches have question marks too.
  - Which flavour of Shrinkage?
  - Hierarchical model fitting issues with overlapping subgroups?
  - Exchangeability assumption unrealistic?
  - Unclear trade-offs by shrinkage when true differential effects are present.
  - Stratified (pre-specified?) approach might be worth considering?
  - Technical model fitting issues sometimes with overlapping subgroups?

- Several **other approaches** suggested ([1], [32, 33], [4]):
  - Some only for selected subgroup? some only for non-overlapping subgroups?
  - Some further head-to-head comparisons of operating characteristics needed?
  - Some question marks on the handling of prognostic factors for some permutation approaches [15]?

Regulator's & other sponsor's view on this entire topic? **What do you think**?

Acknowlegements:

Thanks to many colleagues for various helpful feedback, including:

John Adler, Jonathan Bartlett (former AZ), Anna Berglind, Carl-Fredrik Burman, Mattis Gottlow, Chris Miller and David Wright, with a special thanks for Dan Jackson.

In the event of errors or false claims, the main author takes the full responsibility for it.

References (1).

[1] **Bornkamp, B., Ohlssen D., Magnusson B., Schmidli, H**.; Model Averaging for Treatment effect estimation in subgroups. Phar. Statistics, 2017, vol 16.

[2] **Brookes ST., Whitely E., Egger M., Smith GD., Mulheran PA., Peters TJ.;** *Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test.* J Clin Epidemiol. 2004, Vol 57(3).

[3] **Chen, YH., Wu, YC., Wang M.;** *A Bayesian approach to evaluating regional treatment effect in a multiregional trial*. J. Biopharm Stat., 2009, 19(5).

[4] **Chen J., Zheng H., Quan H., Li G., Gallo P., Ouyang SP., Binkowitz B.**; Ting N, Tanaka Y, Luo X, Ibia E; *Graphical assessment of consistency in treatment effect among countries in multi-regional clinical trials.* Clin Trials. **2013**;10(6).

[5] **Cui L, Hung JHM, Wang SJ, Tsong Y**. *Issues related to subgroup analysis in clinical trials*. *Journal of Biopharmaceutical Statistics* 2002, Vol 12.

[6] **David, C. E., Leffingwell, D. P**.; *Empirical Bayes estimates of subgroup effects in clinical trials*. Controlled Clinical Trials 11, 1990.

[7] **Dane, A., Spencer, A., Rosenkranz, G., Lipkovich, I., Parke, Tom**. *Subgroup Analysis and Interpretation for Phase 3 Confirmatory T7ials: White paper of the EFSPI/PSI Working Group on Subgroup Analysis.* [Submitted 2018], Pharmaceutical Statistics.

[8] **Dane, A., Spencer, A., Stone, A., Svensson, D.**; *SEAMOS: The use of a Resampling Based Graphical methods to Present Standardised Effects Adjusted for Multiple Overlapping Subgroups.* [Submitted 2017] Statistics in Biopharmaceutical Research.

[9] **Dixon D., Simon R.**; *Bayesian Subset Analysis*. Biometrics 47.

References (2).

[10] **Dmitrienko, A., Millen, B., Lipkovich, I.**; *Multiplicity considerations in Subgroup Analysis*. Stat. In Medicine, 2017.

[11] **Efron**, **Bradley.** *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction.* Cambr. Univ. Press 2010.

[12] **Efron**, **Bradley.** *Empirical Bayes Methods for Combining Likelihoods.* 1996, J. Am. Stat. Assoc. 91.

[13] **EMA** *Guideline on the investigation of subgroups in confirmatory clinical trials* http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf).

[14] **Fletcher C, Chuang-Stein C, Paget MA, Reid C, Hawkins N**.; *Subgroup analyses in cost-effectiveness analyses to support health technology assessments.* Pharm.Stat. 2014 Jul-Aug;13(4):265-74

[15] **Foster J., Nan B., Shen L., Kaciroti N., Taylor J.**; *Permutation Testing for Treatment-Covariate Interactions and Subgroup Identification*. Stat. Biosci. 2016 Jun;8(1):77-98

[16] **Gelman A, Carlin J., Stern H.** Bayesian Data Analysis, Chapman Hall, 3rd Ed., 2013.

[17] **Gelman A.** *Prior Distributions for Variance Parameters In Hierarchical Models.* Bayesian Analysis Vol 1(3), 2006.

[18] **Gelman A., & The Stan Development Team**; STAN software: http://mc-stan.org/

[19] **Harpaz R., DuMouchel W., LePendu P., Bauer-Mehren A., Ryan P., Shah N.**; *Performance of Pharmacovigilance Signal Detection Algorithms for the FDA Adverse Event Reporting System*. Clin Pharmacol Ther. 2013 June; 93(6).

References (3).

[20] **Hauck, W.W., Anderson, S., Marcus SM**. *Should we adjust for covariates in nonlinear regression analyses of randomized trials?* Control Clin. Trials 1998, 19(3), 249-56.

[21] **ICH Efficacy Guidelines** *http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html#16*

[22] **Jackson D.,Veroniki A.,Law M.,Tricco A.,Baker R.** *Paule-Mandel estimators for network meta-analysis with random inconsistency effects*. Research Synthesis Methods, Vol 8 (4), 2017.

[23] **Jones H., Ohlssen D., Neuenschwanderc B., Racine A., Branson M.**; *Bayesian models for subgroup analysis in clinical trials.* Clinical Trials 2011; 8: 129–143.

[24] **Lipkovich I. , Dmitrienko A., Denne J., Enas G.**; *Subgroup Identification based on Differential Effect Search (SIDES) – a recursive partitioning method for establishing response to treatment in patient subpopulations*. Stat. In. Medicine, 2011.

[25] **Marius T., Borncamp, B.**; *Comparing Approaches to Treatment Effects Estimation for Subgroups In Clinical Trials*. Stat. in Bioph. Research, 9:2, 2016.

[26] **Parametric Empirical Bayes Inference: Theory and Applications.** Morris C., Journal of the American Statistical Association, Vol. 78, No. 381 (Mar., 1983)

[27] **Miller B., Dmitrienko A., Song G.**; *Bayesian Assessment ot the Influence and Interaction Conditions in Multipopulation Tailoring Clinical Trials*. J. of Biopharm. Stat., Vol 24, 2014.

[28] **Normand T**.; *Tutorial in Biostatistics. Meta-analysis: formulating, evaluating, combining, and reporting by S-L*. Statistics in Medicine, 18 (1999).

References (4).

[29] **Paget M., Chuang-Stein C., Fletcher C., Reid C.;** *Subgroup analyses of clinical effectiveness to support health technology assessments.* Pharm. Stat., Vol 10 (6) 2011.

[30] **Quan H., Li M., Shih WJ., Ouyang, SP., Chen J., Zhang J., Zhao PJ.**; *Empirical shrinkage estimator for consistency assessment of treatment effects in multi-regional clinical trials.* Stat. In Med., Vol 32, Issue 10, 2013.

[31] **Quan H., Li M., Shih WJ., Ouyang SP., Chen J., Zhang J., Zhao PJ.**; *Multi-regional clinical trial design and consistency assessment of treatment effects.* Stat. In Med., Vol 33, Issue 13, 2014.

[32] **Rosenkranz G.;** *Exploratory Subgroup Analyses in Clinical Trials.* ISCB Pre-Conference Training Course, Birmingham, August 21, 2016

[33] **Rosenkranz G.;** *Bootstrap corrections of treatment effect estimates following selection.* Computational Statistics and Data Analysis, Vol 69, 2014.

[34] **Russek-Cohen., E**.; http://www.ema.europa.eu/docs/en_GB/document_library/Presentation/2015/03/WC500183614.pdf., *Comments from the FDA working group on SUBGROUPS*, 2015.

[35] **Varadhan R., Wang SJ.**; Treatment effect heterogeneity for univariate subgroup in clinical trials: Shrinkage, standardization, or else. Biometrical Journal 58, 2016.

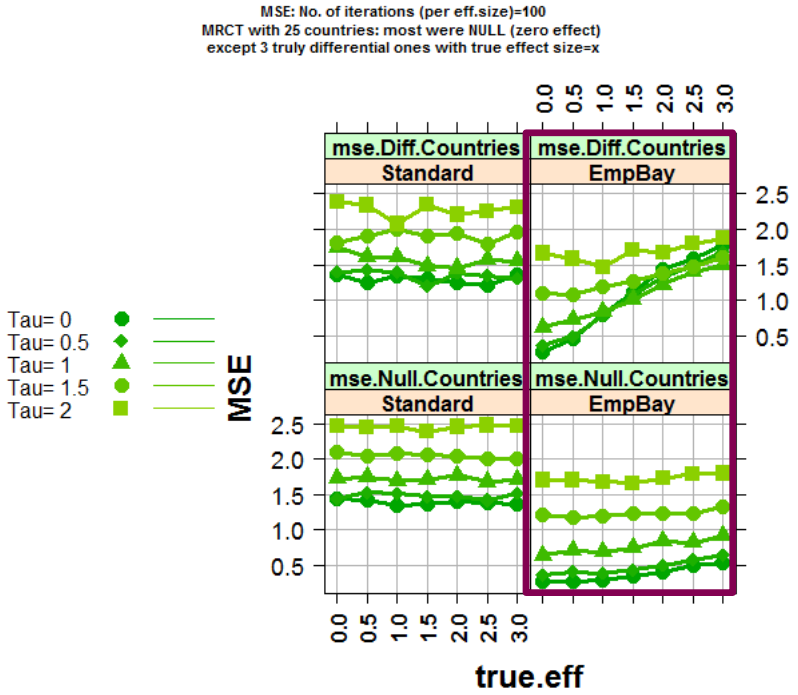
[36] **Stein C.**, *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Berkeley and Los Angeles, University of California Press, 1956, Vol. 1.

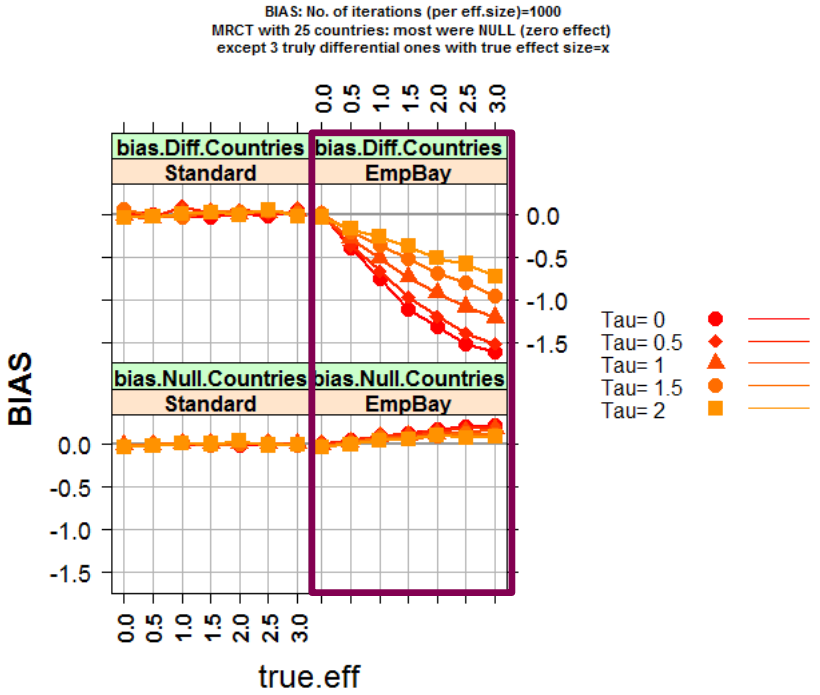
[37] **Viechtbauer W**., METAFOR R package, https://cran.r-project.org/web/packages/metafor/index.html, 2017.

Back-Up Slides:

MSE & Bias (exchangability violated).

X-axis= True Effect (x) in the differential countries.





Confidentiality Notice